Physics-Based Visual Inference: Theory and Applications

Physics-Based Visual Inference: Theory and Algorithms

A DISSERTATION PRESENTED
BY
YING XIONG
TO
SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
ENGINEERING SCIENCES

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
AUGUST 2015

Dissertation advisor: Professor Todd Zickler Ying Xiong

Physics-Based Visual Inference: Theory and Algorithms

Abstract

Analyzing images to infer physical scene properties is a fundamental task in computer vision. It is by nature an ill-posed inverse problem, because imaging is a complicated, information-lossy physical and measurement process that cannot be deterministically inverted. This dissertation presents theory and algorithms for handling ambiguities in a variety of low-level vision problems. They are based on two key ideas: (1) explicitly modeling and reporting uncertainties are beneficial to visual inference; and (2) using local models can significantly reduce ambiguities that would exist in pixelwise analysis.

In the first part of the dissertation, we study the color measurement pipeline of consumer digital cameras, and consider the inherent uncertainty of undoing the effects of tone-mapping. We introduce statistical models for this uncertainty and algorithms for fitting it to given cameras or imaging pipelines. Once fit, the model provides for each tone-mapped color a probability distribution over linear scene colors that could have induced it, which is demonstrated to be useful for a number of downstream inference applications.

In the second part of the dissertation, we study the pixelwise ambiguities in physics-based visual inference and present theory and algorithms for employing local models to eliminate or reduce these ambiguities. In shape from shading, we perform mathematical analysis showing that when restricted with quadratic local models, the shape and lighting ambiguities will be reduced to a small finite number of choices as opposed to otherwise continuous manifolds. We propose a framework for surface reconstruction by enforcing consensus on the local regions, which is later enhanced and extended to be applicable to a variety of other visual inference tasks.

Contents

Listing of figures

Dedicated to *Weiyi*.

Acknowledgments

First of all, I would like to express my deepest appreciation to my advisor, Prof. Todd Zickler, without whom this dissertation would not exist. Todd is an inspiring mentor, a dedicated teacher, and a trusted friend. His insights and thoughts provided essential guidance over the course of my study, and his tremendous support and encouragement helped me overcome every difficulty encountered during my Ph.D. years.

I would like to thank my other committee members, Prof. Steven J. Gortler and Prof. Yue M. Lu for their insightful comments and feedback, which this dissertation has significantly benefited from.

I am fortunate to have the opportunity to collaborate with a lot of wonderful people over the course of my Ph.D. I want to thank Prof. Ronen Basri, Prof. Ayan Chakrabarti, Dr. Manmohan Chandraker, Prof. Trevor Darrell, Dr. Daniel Glasner, Prof. David W. Jacobs, Prof. Anat Levin, Prof. Daniel Scharstein, Prof. Kate Saenko, and Baochen Sun. Working with them has shaped the way I do research. In particular, I would like to express a special gratitude to Prof. Ayan Chakrabarti, with whom I closely worked together on several different projects, and from whom I learned how to tackle the problems and find a way out when getting stuck.

I am always grateful to my parents and family for their tireless support on every aspect of my life. They provided me the education that taught me not only how to study and work, but more importantly, how to be a confident, honest and kind man.

Finally, I dedicate the greatest thanks to my wife, Weiyi, who is the source of all my happiness, passion and love. She is my reason and purpose in writing this dissertation.

1

Introduction

Imaging is a complicated physical and measurement process: light emitted from the source interacts with objects according to their geometry and material properties, gets absorbed, reflected or refracted, and some reaches the capturing device (*e.g.* a camera), which is itself a complex system that performs a number of processing steps before converting the count of photons into final recorded measurements (*e.g.* an sRGB image). In a very general sense, the process can be written as the following equation:

$$image = f_{\text{camera}}\big(f_{\text{interaction}}(illumination,\ material,\ geometry)\big) \tag{1.1}$$

The inference of scene properties from recorded images is an inverse problem in which the $image$ is given as input and physical properties such as $illumination$, $material$ and/or $geometry$ are the

output. This inverse problem is almost always ill-posed, because the imaging processes f_{camera} and $f_{\text{interaction}}$ have information loss. In other words, for a given image, there could be more than one—in fact, usually a large number of—scene combinations that can explain the input equally well. Therefore, a reasonable inference theory or algorithm needs to account for this ambiguity, for example, by exploiting prior knowledge of the physical properties, and/or by reporting the uncertainties to downstream applications (which potentially have more information to make better judgements).

The scope of this dissertation is in physics-based visual inference, also known as "low-level computer vision". We will focus on explicitly modeling the physical or measurement process of image formation, including light-objection interactions and color processing pipeline inside consumer digital cameras. The outputs of our methods are "property maps" that have the same dimension as the input images and try to explain the formation of each individual pixel, *e.g*. a normal vector map in shape from shading, a disparity map for binary stereo, or a linear color image (with uncertainty) for inverse tone-mapping. This is in contrast to "high-level computer vision" approaches which have different goals such as object detection or scene recognition and are generally ignorant or invariant to the exact details of physical process that creates the image.

Low-level vision has a wide range of applications by itself, and can also support and improve high-level vision. Reliably recovering linear scene colors from sRGB images facilitates accurate appearance models for recognition (*e.g*. plant species identification by color), improves radiometric reasoning for tasks like shadow and/or glare removal, and benefits applications such as three-dimensional (3D) reconstruction and virtual tourism that rely on matching and extracting photometric signals. Estimating the depth and/or motion from images (shape from shading, binocular stereo and optical flow) significantly enhances the capabilities of artificial intelligent systems such as robotic navigation and manipulation, autonomous or assisted driving and abnormal events detection. Many algorithms in physics-based visual inference are inspired by biological visual systems, and studying these algorithms might in turn improve our understanding to biological vision, such as human visual perception.

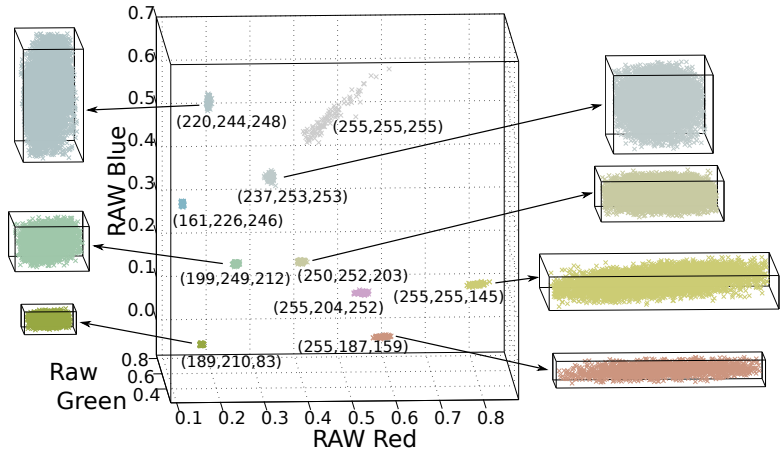


Figure 1.1: Clusters of linear color (RAW) measurements that each map to a single sRGB color value (indicated in parentheses) in a digital SLR camera (Canon EOS 40D). Close-ups of the clusters emphasize the variations in cluster size and orientation. When inverting the tone-mapping process, this is structured uncertainty that cannot be avoided.

1.1 Information Loss in Camera Color Processing

The first part of this dissertation will examine the color processing pipeline of consumer digital cameras. This process, denoted as the $f_{\text{camera}}(\cdot)$ function in Equation (1.1), is the last step of the image formation and therefore the first task for the inverse problem of inference. As will be detailed in Chapter 2 (also see Figure 1.1), consumer digital cameras impose significant distortion and compression to the incoming signals received by their sensors in order to produce a compact and visually pleasing sRGB image for storage and display. The quantization during the process makes the inversion from a compact sRGB image back to linear scene colors impossible, and in this sense information is lost.

This information loss is not always a big problem. For example, in high-level vision tasks such as object detection and recognition, the variation in pose and/or category of the object is more significant than the distortion imposed by the camera color pipeline itself, and the algorithms for such tasks are intentionally designed to be invariant to all such differences in order to extract the higher-level abstraction. Also, some applications such as robotics can afford a more dedicate and better controlled camera that can minimize or even eliminate the distortion, and therefore avoid the problem itself. However,

for low-level vision tasks that use images produced by consumer digital cameras and/or shared on the internet to study the scene properties by explicitly reasoning about the image formation process, the distortion on the camera side is important and worth accounting for.

In this dissertation, we will analyze typical consumer color rendering processing pipelines to understand the source and reason of this distortion and information loss, and present a probabilistic algorithm that undoes such distortion and reports the associated uncertainty. We will show that physics-based visual inference can substantially benefit from such a probabilistic de-rendering step when working on compact sRGB images produced by consumer digital cameras. More details can be found in Chapter 2 of this dissertation.

1.2 Pixelwise Ambiguity and Inference with Local Regions

In the second part of the dissertation, we will focus on visual inference regarding the physical interaction of light and objects in the scene, *i.e.* the $f_{\text{interaction}}(\cdot, \cdot, \cdot)$ function in Equation (1.1), assuming the effect of $f_{\text{camera}}(\cdot)$ function has already been inverted with the uncertainties properly accounted for. The key observation in this part of the dissertation is that physics-based inference is usually pixelwise ambiguous and it is usually beneficial to perform such inference on regions of appropriate sizes. More specifically, in most scenarios, it is not possible to infer the unique scene property from a single pixel measurement of the input image, because multiple different properties can explain the pixel measurement equally well. However, we will present theory and algorithms showing that by considering a bigger neighborhood of image data, the ambiguity can be significantly reduced by applying proper local models, and furthermore, the remaining ambiguity can be explicitly characterized and used in a global reasoning framework to accurately recover the entire scene property map.

One physics-based inference problem we consider is Lambertian shape from shading, in which the intensity of a pixel is assumed to be the dot product of directional lighting and surface normal: $I = n \cdot l$. It is easy to see that looking at the intensity of a single pixel I, even with a known lighting

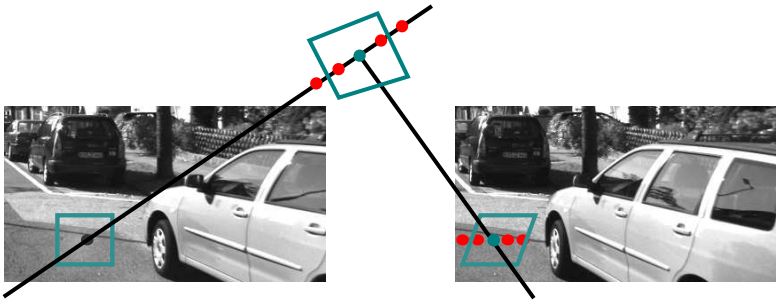


Figure 1.2: Pixelwise stereo matching is ambiguous because the pixel in one view can potentially match many pixels equally well on the other view. Performing the match by local regions will significantly reduce such ambiguity, as long as the regions are ``of the right size'' that contain enough contextual information but does not cross disparity boundary.

l and in the absence of noise, the normal vector n can still not be uniquely determined and possible values form a one-dimensional cone around l (or a conic curve on on the projective plane). In Chapter 3 of the dissertation, we prove that by looking at a larger image region (say a 5×5 square patch) and assuming a quadratic local model, the lighting and surface can be tightly restricted, say to a four-way choice instead of a continuous 5D manifold. We also analyze the possible implication of noise in the inference process, and show that the possible shapes can still be effectively restricted to a low-dimensional manifold when considering modest amount of noise.

Another inference problem discussed in this dissertation is binocular stereo matching. As shown in Figure 1.2, trying to match the image intensities pixelwise from one view to another is usually noisy and unreliable, as pixel in one image can potentially have many possible good matches on the other view. Performing the matching with local regions is a common technique used to reduce such ambiguity, and one needs to select the regions of the right size such that they contain enough contextual information but also are not too big as to cross disparity boundaries. In Chapter 4, we introduce a multi-scale framework that simultaneously decide which regions are of the right size for inference, and for regions of the right size, find their best local model parameters.

1.3 Citations to Previously Published Work

The dissertation is organized such that each chapter contains a "related work" section describing relevent references to that chapter. In this section, we list a few citations to my own published work that compose most of this dissertation.

Most of Chapter 2 has been published as

- Ying Xiong, Kate Saenko, Travor Darrell and Todd Zickler. "From pixels to physics: Probabilistic color de-rendering". *Computer Vision and Pattern Recognition, IEEE Conference on,* 2012.

- Ayan Chakrabarti, Ying Xiong, Baochen Sun, Trevor Darrell, Daniel Scharstein, Todd Zickler and Kate Saenko. "Modeling radiometric uncertainty for vision with tone-mapped color images". *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 2014.

Most of Chapter 3 has been published as

- Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs and Todd Zickler. "From shading to local shape". *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 2015.

A significant portion of Chapter 4 has been published as

- Ayan Chakrabarti, Ying Xiong, Steven J. Gortler, Todd Zickler. "Low-level Vision by Consensus in a Spatial Hierarchy of Regions". *Computer Vision and Pattern Recognition, IEEE Conference on,* 2015.

Note that some of the work is in collaboration with Dr. Ayan Chakrabarti. This dissertation mostly contains contributions made primarily by myself. Contributions lead by Dr. Ayan Chakrabarti are also included but described in less detail, particularly, in Section 2.3.3, Section 3.5.2 and Section 4.5.

2

Probabilistic Color De-rendering

2.1 Introduction

Most digital images produced by consumer cameras and shared online exist in narrow-gamut, low-dynamic range formats.[1] This is efficient for storage, transmission, and display, but it is unfortunate for computer vision systems that seek to interpret this data radiometrically when learning object appearance models for recognition, reconstructing scene models for virtual tourism, or performing other visual tasks with Internet images. Indeed, most computer vision algorithms are based, either implicitly or explicitly, on the assumption that image measurements are proportional to the spectral radiance of the scene (called *scene color* hereafter), and when a consumer camera renders its digital linear color mea-

[1] Typically sRGB color space with JPEG encoding: IEC 10918-1:1994 and IEC 61966-2-1:1999

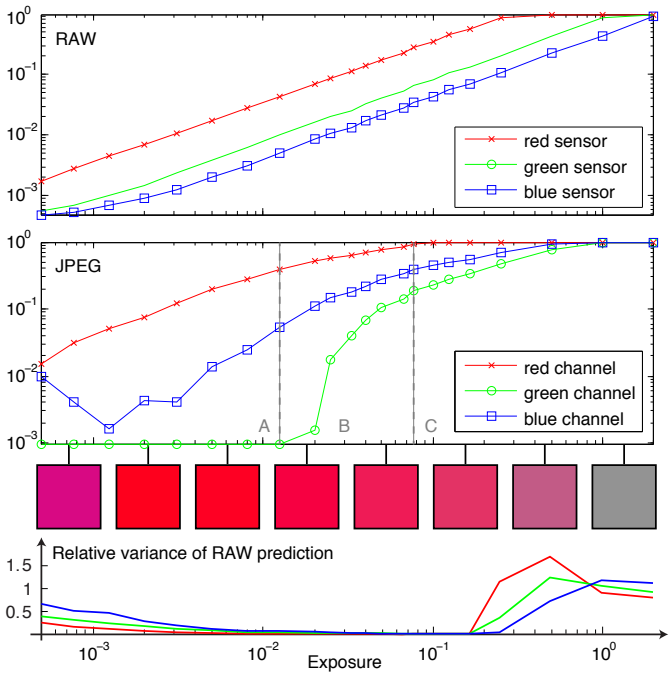


Figure 2.1: RAW and JPEG values for different exposures of the same spectral scene radiance collected by a consumer digital camera (DMC-LX3, Panasonic Inc.), along with normalized-RGB visualizations of the reported JPEG colors at a subset of exposures. Apart from sensor saturation, RAW values are linear in exposure and proportional to spectral irradiance; but narrow-gamut JPEG values are severely distorted by tone-mapping. Given only JPEG values, what can we say about the unknown RAW values---and thus the scene color---that induced it? How can we use all of the JPEG color information, including when some JPEG channels are saturated (regions A and C)? We answer these questions by providing a confidence level for each RAW estimate (bottom plot), which can benefit radiometry-based computer vision.

surements to a narrow-gamut output color space (called *rendered color* hereafter), this proportionality is almost always destroyed. Figure 2.1 shows an example.

Existing approaches to color de-rendering attempt to undo the effects of a camera's color processing pipeline through "radiometric calibration" [13, 51, 56], in which rendered colors (*i.e.*, those reported in a camera's JPEG output) are reverse-mapped to corresponding scene colors (*i.e.*, those that would have been reported by the same camera's RAW output) using a learned deterministic function. This approach is unreliable, because it ignores the inherent uncertainty caused by the loss of information. A typical camera renders many distinct sensor measurements to the same small neighborhood of narrow-

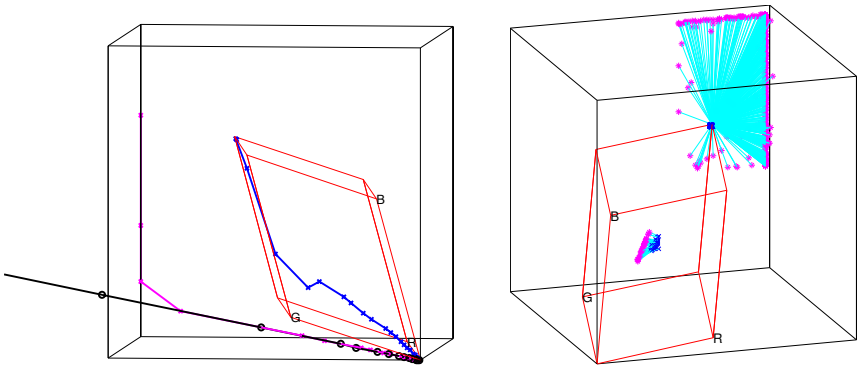


Figure 2.2: 3D visualization of color rendering. The black cube indicates the set of possible RAW color sensor measurements, and the red parallelepiped shows the boundary of the output sRGB gamut to which all RAW colors must be tone-mapped.[2] *Left*: Data from Figure 2.1, with black circles the scene color $\mathbf{x}$ at different exposure times. Corresponding RAW values $\tilde{\mathbf{x}}$ (magenta) are clipped due to sensor saturation, and they are tone-mapped to rendered colors $\mathbf{y}$ (blue) within the output sRGB gamut. *Right*: Rendered colors (blue) in small neighborhoods of $(127, 127, 127)$ and $(253, 253, 253)$ in a JPEG image, connected (through cyan lines) to their corresponding RAW measurements (magenta).

gamut output colors (see Figure 2.2, right) and, once these output colors are quantized, the reverse mapping becomes one-to-many in some regions and cannot be deterministically undone.

How can we know which predictions are unreliable? As supported by Figure 2.2, one expects the one-to-many effect to be greatest near the edges of the output gamut (*i.e.*, near zero or 255 in an 8-bit JPEG file), and practitioners try to mitigate it using heuristics such as ignoring all JPEG pixels having values above or below certain thresholds in one or more of their channels. This trick improves the reliability of deterministic radiometric calibration, but it raises the question of how to choose thresholds for a given camera. ("Should I only discard pixels with values 0 or 255, or should I be more conservative?")[3] A more fundamental concern is that this heuristic works by discarding information that would otherwise be useful. Referring to Figure 2.1, such a heuristic would ignore all JPEG measure-

[2] The boundary of the output sRGB gamut is determined automatically from image data in two steps. The edge directions of the parallelepiped are extracted from RAW metadata using dcraw[12], and then its scale is computed as a robust fit to RAW-JPEG correspondences.

[3] Our experiments in Figure 2.1 and those of [49] reveal significant variation between models and suggest the answer is often the latter.

ments in regions A and C, even though these clearly tell us *something* about the latent scene color.

To overcome these limitations, we introduce two probabilistic approaches for de-rendering. These methods produce from each rendered (JPEG) color a probability distribution over the (wide gamut, high dynamic range) scene colors that could have induced it. They rely on an offline calibration procedure involving registered RAW and JPEG image pairs, and from these infer a statistical relationship between rendered colors and scene colors. The first approach uses a local Gaussian Process (GP) to perform regression from rendered colors to scene colors. The second approach, developed in a collaboration work lead by Ayan Chakrabarti [9], learns a deterministic forward mapping from scene colors to rendered colors and then does a probabilistic inverse by analyzing the conditional distribution of scene colors given a rendered color based on learned forward mapping. Both approaches provide a measure of confidence, based on the variance of the output distribution, for every predicted scene color, thereby eliminating the need for heuristic thresholds and making better use of the scene radiance information that is embedded in an Internet image. The offline calibration procedure is required only once for each different imaging mode of each camera, thus many per-camera de-rendering models could be stored in an online database and accessed on demand using camera model and mode information embedded in the metadata of an Internet image.[4]

We evaluate our approach in a few different ways. First, we assess our ability to recover wide-gamut scene colors from JPEG sRGB observations in different consumer cameras. Next, we employ our probabilistic de-rendering model in relatively straightforward probabilistic adaptations of two established applications: high-dynamic range imaging with an exposure-stack of images (*e.g.*, [51]) and three-dimensional reconstruction via Lambertian photometric stereo (*e.g.*, [85]). In all cases, a probabilistic approach significantly improves our ability to infer radiometric scene structure from tone-mapped images.

[4] As has been done for lens distortion by PTLens (accessed Mar 27, 2012): http://www.epaperpress.com/ptlens/

2.2 Related Work

The problem of radiometric calibration, where the goal is inverting non-linear distortions of scene radiance that occur during image capture and rendering, has received considerable attention in computer vision. Until recently, this calibration has been formulated only for grayscale images, or for color images on a per-channel-basis by assuming that the "radiometric response function" in each channel acts independently [51, 56, 13, 23]. While early variants of this approach parametrized these response functions simply as an exponentiation (or "gamma correction") with the exponent as a single model parameter, later work sought to improve modeling accuracy by considering more general polynomial forms [23]. Since these models have a relatively small number of parameters, they have featured in several algorithms for "self-calibration"—parameter estimation from images captured in the wild, without calibration targets—through analysis of edge profiles [50, 78], image statistics [18, 45], or exposure stacks of images [51, 56, 13, 22, 69, 73].

However, per-channel models cannot accurately model the color processing pipelines of most consumer cameras, where the linear sensor measurements span a much wider gamut than the target output format. To be able to generate images that "look good" on limited-gamut displays, these cameras compress out-of-gamut and high-luminance colors in ways that are as pleasing as possible, for example by preserving hue. This means that two scene colors with the same raw sensor value in their red channels can have very different red values in their mapped JPEG output if one RAW color is significantly more saturated than the other.

Chakrabarti et al. [8] investigated the accuracy of more general, cross-channel parametric forms for global tone-mapping in a number of consumer cameras, including multi-variate polynomials and combinations of cross-channel linear transforms with per-channel polynomials. While they found reasonable fits for most cameras, the residual errors remained relatively high even though the calibration and evaluation were both limited to images of a single relatively narrow-gamut chart. Kim et

al. [39] improved on this by explicitly reasoning about the mapping of out-of-gamut colors. Their model consists of a cascade of: a linear transform, a per-channel polynomial, and a cross-channel correction for out-of-gamut colors using radial basis functions. The forward tone-map model we use in Section 2.3.3 is strongly motivated by this work, although we find a need to augment the calibration training data so that it better covers the full space of measurable RAW values.

All of these approaches are focussed on modeling the distortion introduced by global tone-mapping. They do not, however, consider the associated loss of information, nor the structured uncertainty that exists when the distortion is undone as a pre-process for radiometric reasoning by vision systems. Indeed, while the benefit of undoing radiometric distortion has been discussed in the context of various vision applications (*e.g.*, deblurring [10, 78], high-dynamic range imaging [60], video segmentation [24]), previous methods have relied exclusively on deterministic inverse tone-maps that ignore the structured uncertainty evident in Figures 2.2. The main goal of this of this paper is to demonstrate that the benefits of undoing radiometric distortion can be made significantly greater by explicitly modeling the uncertainty inherent to inverse tone-mapping, and by propagating this uncertainty to subsequent visual inference algorithms.

Finally, we note that our proposed framework applies to stationary, global tone-mapping processes, meaning those that operate on each pixel independent of its neighboring pixels, and are unchanging from scene to scene. This is applicable to many existing consumer cameras locked into fixed imaging modes ("portrait", "landscape" *etc.*), but not to local tone-mapping operators that are commonly used for HDR tone-mapping.

2.3 Probabilistic De-rendering Models

We begin with a model for the forward color processing pipeline of a typical consumer digital camera; then we describe two approaches to represent and fit the reverse mapping. The models in this section ignore secondary effects such as de-mosaicking, flare removal, noise removal, sharpening, *etc.*, since

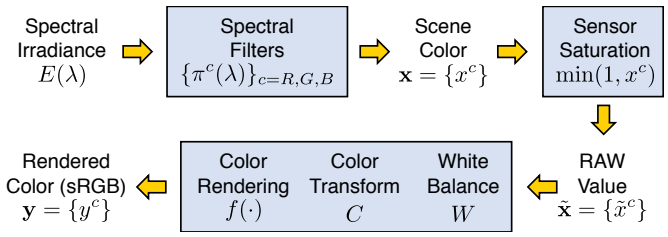


Figure 2.3: The forward color processing model used in this chapter, along with our notation for it. Lesser effects, such as flare removal, de-mosaicking, and vignetting are ignored and treated as noise.

these have significantly less impact on the output than non-linear tone-mapping. More details on these issues can be found elsewhere [8, 5, 65, 31].

An important assumption underlying our model is that the forward rendering operation is spatially-uniform, meaning that its effect on a RAW color vector is the same regardless of where it occurs on the image plane. This assumption is shared by almost all de-rendering techniques and is reasonable at present; but if spatially-varying tone-mapping operators become more common, relaxing this assumption may become a useful direction for future work.

2.3.1 Forward (Rendering) Model

Referring to Figure 2.3, the forward model begins with three idealized spectral sensors with sensitivity profiles $\{\pi^c(\lambda)\}_{c=R,G,B}$ that sample the spectral irradiance incident on the sensor plane. These sensors are idealized in that they do not saturate and have infinite dynamic range, and we refer to their output $\mathbf{x} = \{x^c\}_{c=R,G,B}$ as the *scene color*. Practical sensors have limited dynamic range, so scene colors are clipped as they are recorded. In some consumer cameras these recorded sensor measurements $\tilde{\mathbf{x}} = \{\tilde{x}^c\}_{c=R,G,B}$ are made available through a RAW output format, and in others they only exist internally. Empirical studies suggest that the RAW values (in the absence of clipping) are proportional to incident irradiance and related by a linear transform to measurements that would be obtained by the CIE standard observer [8, 5, 38] (also see Figure 2.1). For this reason, they provide

a "relative scene-referred image" [27] and can be used directly by computer vision systems to reason about spectral irradiance.

Two linear transforms are applied to the sensor measurements. The first (W) is scene-dependent and induces white balance, and the second (C) is a fixed transformation to an internal working color space. Then, most importantly, the linearly transformed RAW values $CW\tilde{\mathbf{x}}$ are rendered to colors $\mathbf{y} = \{y^c\}_{c=R,G,B}$ in the narrow-gamut output sRGB color space through a non-linear map f: $\mathbb{R}^3 \rightarrow \mathbb{R}^3$. This map has evolved to produce visually-pleasing results at the expense of physical accuracy, and since the quality of a camera's color rendering process plays a significant role in determining its commercial value, there is a dis-incentive for manufacturers to share its details. In our model, the map f includes the per-channel non-linearity that is part of the sRGB standard (IEC 61966-2-1:1999).

The left of Figure 2.2 shows signal values at various stages of this forward model for a consumer camera (DMC-LX3, Panasonic Inc.). Recall that the black box in this plot represents the range of possible RAW values $\tilde{\mathbf{x}}$, and the red parallelepiped marks the boundary of the output sRGB gamut. The plot shows color signals produced using different exposure times for a simple static scene consisting of a uniform planar patch under constant illumination, with spatial-averaging over all patch pixels to thoroughly suppress the effects of noise, demosaicking, and JPEG compression. The scene colors $\mathbf{x}$ (black) lie a line that extends well beyond the cube as the exposure time grows large, and the chromaticity of the patch is such that all scene colors lie outside the sRGB gamut. The wide-gamut RAW values $\tilde{\mathbf{x}}$ (magenta) are very close to these scene colors for low exposures, but they are clipped for longer exposures when the intensity grows large. The rendered colors $\mathbf{y} = f(CW\tilde{\mathbf{x}})$ (blue) lie within the output gamut, and are significantly affected by the combined effects of sensor saturation, white balance, and the color space transform. Interestingly, these rendered colors are relatively far inside the boundary of the sRGB gamut, so the conventional wisdom in radiometric calibration that one should discard pixels with very small or very large JPEG values as being "clipped" is unlikely to detect and properly treat them.

2.3.2 Inverse (De-rendering) Model with Gaussian Process

Our goal is to infer, for each possible rendered color $\mathbf{y}$, the original scene color $\mathbf{x}$ that created it. As information is lost in the forward rendering process, exact recovery is not possible and thus any deterministic function that predicts a single point estimate is bound to be wrong much of the time. For that reason, we propose to estimate a *distribution* over the space of possible scene colors. Specifically, we seek a representation of $p(\mathbf{x}|\mathbf{y})$ from which we can either obtain a MAP estimate of $\mathbf{x}$ or directly employ Bayesian inference as desired for a given application (see Section 2.4.1 and Section 2.4.2).

We model the underlying de-rendering function, denoted z, using Gaussian process (GP) regression [68]. Given a training set $\{\mathcal{D} = (\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \cdots, N\}$, composed of inputs $\mathbf{y}_i$ and noisy outputs $\mathbf{x}_i$, we model the outputs $\{x_i^c\}_{c=R,G,B}$ in each channel separately as coming from a latent function z^c that has a prior distribution described by a GP, and is corrupted by additive noise ϵ_i:

$$x_i^c = z^c(\mathbf{y}_i) + \epsilon_i, \quad \epsilon_i \propto \mathcal{N}(0, \sigma_n^2). \tag{2.1}$$

The latent function z serves as the inverse of the forward rendering map composed of the color rendering function, color transform, and white balance operations depicted in Figure 2.3. We will learn it using images in which the white balance has been fixed to remove scene-dependence.

The classic GP regression paradigm uses a single set of hyper-parameters controlling the smoothness of the inferred function. However, our analysis of camera data has revealed that such globally-defined (*i.e.*, stationary) smoothness is inadequate because there is significantly different behavior in different regions of the sRGB gamut (see right of Figure 2.2.) Instead, the variance of z should be allowed to vary over local neighborhoods of the sRGB color space.

Several extensions to the classic GP have been proposed to model input-varying noise [67, 80, 54]. Here, we employ a *local* GP regression model, which exploits the observation that, for compact radial covariance functions, only the points close to a test point have significant influence on the results [80].

Given a training dataset and a test point, the method identifies a local neighborhood of the test point, and performs prediction with the model either pre-trained based on some local cluster ("offline local GP"), or learned on the fly using neighbor points just detected ("online local GP").[5] More precisely, given training set $\mathcal{D}$ and a test sRGB color $\mathbf{y}$, we infer a test distribution of RAW values $\mathbf{x}$ conditioned on $\mathbf{y}$ by identifying a local neighborhood of $\mathbf{y}$ in $\mathcal{D}$, denoted $\mathcal{D}_{N(\mathbf{y})}$, and computing

$$p_x(\mathbf{x}|\mathbf{y}) = \prod_c p_{GP}(x^c|\mathcal{D}_{N(\mathbf{y})}, \mathbf{y}), \tag{2.2}$$

where $p_{GP}(x|D, \mathbf{y})$ is the conditional GP likelihood of x using training data D for sRGB colors $\mathbf{y}$.

2.3.3 Probablistic Inverse of a Learned Forward Mapping

In a more recent collaboration work lead by Ayan Chakrabarti, we present a new de-rendering approach that first learns a forward mapping from rendered colors to scene colors, and then probabilistically invert it by analyzing the conditional distribution of possible scene colors that could have produced a given rendered color. We briefly summarize the approach in this section, and refer interested readers to [9] for more details.

We model the forward mapping $\mathbb{J} : \mathbf{x} \rightarrow \mathbf{y}$ with a two-step approach: (1) a linear transform followed and independent per-channel polynomial; followed by (2) a correction to account for deviations

[5] To handle multimodality in the mapping, [80] shows how clustering may be performed in both input and output spaces for the training data, and a set of local regressors returned. However we believe that our inverse map does not have multimodal structure, and we found that a single local regressor provided adequate results. Implementation details with regard to online and offline models are described in Section 2.5.

in the rendering of saturated and out-of-gamut colors.

$$\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{y}_R \\ \tilde{y}_G \\ \tilde{y}_B \end{bmatrix} = \begin{bmatrix} f(\mathbf{v}_R^T \mathbf{x}) \\ f(\mathbf{v}_G^T \mathbf{x}) \\ f(\mathbf{v}_B^T \mathbf{x}) \end{bmatrix}, \tag{2.3}$$

$$\mathbf{y} = Q\left(B(\tilde{\mathbf{y}}) + \begin{bmatrix} g_1(\tilde{\mathbf{y}}) \\ g_2(\tilde{\mathbf{y}}) \\ g_3(\tilde{\mathbf{y}}) \end{bmatrix} \right), \tag{2.4}$$

where $\mathbf{v}_R, \mathbf{v}_G, \mathbf{v}_B \in \mathbb{R}^3$ define a linear color space transform, $B(\cdot)$ bounds its argument to the range $[0, 255]$, and $Q(\cdot)$ quantizes its arguments to 8-bit integers. The per-channel non-linearity $f(\cdot)$ is modeled with a polynomial of degree d:

$$f(x) = \sum_{i=0}^{d} \alpha_i x^i. \tag{2.5}$$

Motivated by the observations in [39], this polynomial model is augmented with an additive correction function $g(\cdot)$ in (2.4) to account for deviations that result from camera processing to improve the visual appearance of rendered colors. We use support-vector regression (SVR) with a Gaussian radial basis function (RBF) kernel to model these deviations, *i.e.*, each $g_c(\cdot), c \in \{R, G, B\}$ is of the form:

$$g_c(\tilde{y}) = \sum_i \lambda_{c:i} \exp\left(-\gamma_c \|\tilde{\mathbf{y}} - \mathbf{y}_{c:i}\|^2\right). \tag{2.6}$$

Once having the forward model $\mathbb{J}(\cdot)$, we can characterize all scene colors $\mathbf{x}$ that will be rendered into a given color $\mathbf{y}$, *i.e.* all $\mathbf{x}$ such that $\mathbb{J}(\mathbf{x}) = \mathbf{y}$. Considering the fact that $\mathbb{J}(\cdot)$ itself is learned from training data, we add some slack for calibration error and treat the term $\|\mathbf{y} - \mathbb{J}(\mathbf{x})\|$ as Gaussian noise with variance σ_f^2, which leads to a conditional distribution of scene colors $\mathbf{x}$ given a rendered color $\mathbf{y}$:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z}\, p(\mathbf{x}) \exp\left(-\frac{\|\mathbf{y} - \mathbb{J}(\mathbf{x})\|^2}{2\sigma_f^2}\right), \tag{2.7}$$

where Z is the normalization factor

$$Z = \int p(\mathbf{x}') \exp\left(-\frac{\|\mathbf{y} - \mathbb{J}(\mathbf{x}')\|^2}{2\sigma_f^2}\right) d\mathbf{x}', \tag{2.8}$$

and $p(x)$ is a *prior* on sensor-measurements, which is assumed to be uniform over all possible sensor measurements in this chapter.

We can compute the mean and variance of this conditional distribution as

$$\mu(\mathbf{y}) = \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x}, \tag{2.9}$$

$$\Sigma(\mathbf{y}) = \int (\mathbf{x} - \mu(\mathbf{y}))(\mathbf{x} - \mu(\mathbf{y}))^T p(\mathbf{x}|\mathbf{y}) d\mathbf{x}. \tag{2.10}$$

The integrations in are performed numerically, and by storing pre-computed values of $\mathbb{J}$ on a densely-sampled grid to speed up distance computations. Note that here $\mu(\mathbf{y})$, in addition to being the mean of the conditional distribution, is *also* the single best estimate of $\mathbf{x}$ given $\mathbf{y}$ (in the minimum least-squares error sense) from the exact distribution in (2.7). And since (2.7) is derived using a camera model similar to that of [39], $\mu(\mathbf{y})$ can be interpreted as the deterministic RAW estimate that would be yielded by the algorithm in [39].

Furthermore, we can use $(\mu(\mathbf{y}), \Sigma(\mathbf{y}))$ to approximate the conditional distribution of scene colors $\mathbf{x}$ given a rendered color $\mathbf{y}$ as a multi-variate Gaussian, *i.e.*

$$p(\mathbf{x}|\mathbf{y}) \approx \mathcal{N}(\mathbf{x}; \mu(\mathbf{y}), \Sigma(\mathbf{y})). \tag{2.11}$$

In this sense the output by this approach can be thought as a more general version of the local GP

output (2.2) described in the previous section, because the covariance matrix $\Sigma(\mathbf{y})$ is a general 3×3 matrix where as (2.2) implies a diagonal covariance matrix. In the rest of this paper, when describing a general approach using the photometric uncertainty (Section 2.4), we are ignorant of which approach is used to estimate the uncertainty; when evaluating on datasets (Section 2.5), we use the approach described in this section instead of local GP because the general covariance matrix is more flexible in capturing the true conditional distribution of scene colors and therefore produces higher estimation accuracy compared to the diagonal one.

2.4 Working with Photometric Uncertainty

Linear measurements of scene radiance are crucial for many computer vision tasks (shape from shading, image-based rendering, deblurring, color constancy, intrinsic images, *etc*.), and the output of our de-rendering model can be readily used in probabilistic approaches to these tasks. Here we describe two such tasks and show how modeling photometric uncertainty leads to more robust results.

2.4.1 Probabilistic Wide Gamut Imaging

Many applications that use Internet images operate by inferring radiometric scene properties from multiple observations of the same scene point. For example, multiple observations under different illuminations can be exploited for inferring diffuse object color [59] or more general BRDFs [25]. To explore the benefits of modeling photometric uncertainty in such cases, we consider an example scenario motivated by traditional HDR imaging with exposure stacks [51, 13]. Given as input multiple exposures of the same stationary scene, we seek to combine them into one floating-point, HDR, and wide-gamut image.

Assume we are given a sequence of sRGB vectors captured at shutter speeds of $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ seconds. Represent these by $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. We would like to predict the RAW color, $\mathbf{x}_0$ say, that would have been obtained with a shutter speed of α_0 seconds. Note that α_0 need not be one of the

shutter speeds used to capture the sRGB input.

Given a training set $\mathcal{D}$, for each sRGB value $\mathbf{y}$ we estimate the conditional distributions $p_{x_i}(\mathbf{x}_i|\mathbf{y}_i)$ for the RAW value $\mathbf{x}_i$ that would have been obtained with shutter speed α_i. Then, to obtain $\mathbf{x}_0$, we combine them using

$$p_{x_0}(\mathbf{x}_0|\mathbf{y}_1,\ldots,\mathbf{y}_N) = \prod_i p_{x_0}(\mathbf{x}_0|\mathbf{y}_i) = \prod_i \frac{\alpha_i}{\alpha_0} p_{x_i}\left(\frac{\alpha_i}{\alpha_0}\mathbf{x}_0|\mathbf{y}_i\right). \tag{2.12}$$

Since each channel $p_{x_i}(\mathbf{x}_i|\mathbf{y}_i)$ is modeled as a Gaussian distribution, the conditional distribution $p_{x_0}(\mathbf{x}_0|\mathbf{y}_1,\ldots,\mathbf{y}_N) = \prod_i p_{x_0}(\mathbf{x}_0|\mathbf{y}_i)$ will be Gaussian as well. Our output for $\mathbf{x}_0$, therefore, is the mean and variance of this Gaussian distribution.

This application reveals the power of a probabilistic model: it provides a distribution rather than a point estimate. For applications that combine multiple independent measurements, this provides a natural way to assign more weight to the estimates that have smaller variance.

2.4.2 Probabilistic Lambertian Photometric Stereo

When illumination varies, another way that multiple observations of the same scene can be used is to recover lighting information and/or scene geometry. This may be useful when using Internet images for weather recovery [72], geometric camera calibration [46], or 3D reconstruction [1]. To quantitatively assess the utility of uncertainty modeling in these types of applications we consider the toy problem of recovering from JPEG images three-dimensional scene shape using Lambertian photometric stereo.

Lambertian photometric stereo is a technique for estimating the surface normals of a Lambertian object by observing that object under different lighting conditions and a fixed viewpoint [85]. Suppose there are N different directional lighting conditions, with $\mathbf{l}_i \in \mathbb{R}^3$ the direction and strength of the ith source. Consider a single color channel of single pixel in the image plane; denote by I_i the *linear*

intensity recorded in that channel under the ith light direction; and let $\mathbf{n} \in \mathbb{S}^2$ and $\rho \in \mathbb{R}^+$ be the normal direction and the albedo of the surface patch at the back-projection of this pixel. The Lambertian reflectance model provides the relation $\rho\langle \mathbf{l}_i, \mathbf{n}\rangle = I_i$, and the goal of photometric stereo is to infer the material ρ and shape $\mathbf{n}$ given the set $\{\mathbf{l}_i, I_i\}$.

Defining a pseudo-normal $\mathbf{b} \triangleq \rho\mathbf{n}$, the relation between the observed intensity and the scene parameters becomes

$$\mathbf{l}_i^T \mathbf{b} = I_i. \tag{2.13}$$

Given three or more $\{\mathbf{l}_i, I_i\}$-pairs, the traditional Lambertian photometric stereo estimates pseudo-normal $\mathbf{b}$ (and thus ρ and $\mathbf{n}$) in a least-squares sense:

$$\mathbf{b} = (\boldsymbol{L}^T \boldsymbol{L})^{-1} \boldsymbol{L}^T \mathbf{I}, \tag{2.14}$$

where $\boldsymbol{L}$ and $\mathbf{I}$ are the matrix and vector formed by stacking the light directions $\mathbf{l}_i$ and measurements I_i, respectively.

The linear relation between intensity $\mathbf{I}$ and scene radiance is crucial in photometric stereo. One can use RAW measurements when they are available, but for Internet-based vision tasks that rely on sRGB images, one must first de-render the colors to achieve this linearity. In our case, the de-rendering result for each pixel is described as a Gaussian random variable $I_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and Eq. (2.13) can be re-written as

$$\mathbf{l}_i^T \mathbf{b} = \mu_i + \sigma_i \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1). \tag{2.15}$$

From this it follows (*e.g.*, [29]) that the maximum likelihood estimate of the pseudo-normal $\mathbf{b}$ is obtained through weighted least-squares, with weights given by the reciprocal of the variance. That is,

$$\mathbf{b} = (\boldsymbol{L}^T \boldsymbol{W} \boldsymbol{L})^{-1} \boldsymbol{L}^T \boldsymbol{W} \boldsymbol{\mu}, \quad \text{with } \boldsymbol{W} = \text{diag}\{\sigma_i^{-2}\}_{i=1}^N. \tag{2.16}$$

Once again we see that distributions provided by a probabilistic de-rendering system can be employed very naturally to selectively weight measurements for improved accuracy and robustness.

2.5 Evaluation

Our database consists of images captured using a number of popular consumer cameras, using an X-Rite 140-patch color checker chart as the calibration target as in [8] and [39]. However, although the chart contains a reasonably wide gamut of colors, these colors only span a part of the space of possible RAW values that can be measured by a camera sensor. To be able to reliably fit the behavior of each camera's tone-mapping function in the full space of measurable scene colors, and to accurately evaluate the quality of these fits, we captured images of the chart under *sixteen* different illuminants (we used a standard Tungsten bulb paired with different commercially available gel-based color filters) to obtain a significantly wider gamut of colors. Moreover, for each illuminant, we captured images with different exposure values that range from one where almost all patches are under-exposed to one where all are over-exposed. We expect this collection of images to represent an exhaustive set that includes the full gamut of irradiances likely to be present in a scene.

Most of the cameras in our dataset allow access to the RAW sensor measurements, and therefore directly give us a set of RAW-JPEG pairs for training and evaluation. For JPEG-only cameras, we captured a corresponding set of images using a RAW-capable camera. To use the RAW values from the second camera as a valid proxy, we had to account for the fact that the exposure steps in the two cameras were differently scaled (but available from the image metadata), and for the possibility that the RAW proxy values in some cases may be clipped while those recorded by the JPEG camera's sensors were not. Therefore, the exposure stack for each patch under each illuminant from the RAW camera was used to estimate the underlying scene color at a canonical exposure value, and these were then mapped to the exposure values from the JPEG camera without clipping.

Camera Name	Deterministic Inverse Uniform 8k samples	Prob. Inverse Uniform 8k samples	Prob. Inverse 10 Exp., 2 Illum.	Prob. Inverse 4 Exp., 4 Illum.	Prob. Inverse 8 Exp., 4 Illum.
Panasonic DMC LX3	3.50	12.44	6.19	11.87	12.17
Canon EOS 40D	3.45	13.06	-0.18	11.87	12.22
Canon PowerShot G9	2.01	8.33	7.12	7.80	8.16
Canon PowerShot S90	3.83	11.34	10.47	10.96	10.91
Nikon D7000	1.59	8.52	6.20	3.45	8.28

Table 2.1: Mean Empirical log-Likelihoods under Inverse Models for RAW-capable Cameras.

Camera Name	Deterministic Inverse	Prob. Inverse
Fujifilm J10	1.97	8.69
Panasonic DMC LZ8	1.60	11.83
Samsung Galaxy S3	2.23	7.51

Table 2.2: Mean Empirical log-Likelihoods for JPEG-only Cameras.

2.5.1 De-rendering

To begin, we demonstrate the benefit of using probabilistic de-rendering to hallucinate scene colors from a single narrow-gamut sRGB image. We report the mean empirical log-likelihood, *i.e.*, the mean value of $\log p(\mathbf{x}|\mathbf{y})$ across all RAW-JPEG pairs $(\mathbf{x}, \mathbf{y})$ in the validation set, for our set of calibrated cameras. For comparison, the log-likelihood scores from a deterministic inverse that outputs single prediction ($\mu(\mathbf{y})$ from (2.9)) for the RAW value for a given JPEG is also reported. Note that strictly speaking, the log-likelihood in this case would be $-\infty$ unless $\mu(\mathbf{y})$ is exactly equal to $\mathbf{x}$. The scores reported in Tables 2.1 and 2.2 are therefore computed by using a Gaussian distribution with variance equal to the mean prediction error (which is the choice that yields the maximum mean log-likelihood). We find that these scores are much lower than those from the probabilistic model, demonstrating its benefits in the de-rendering task.

For RAW-capable cameras, we also experimented our de-rendering model using different subsets of collected RAW-JPEG pairs. The first of these subsets is simply constructed with 8000 random RAW-JPEG pairs sampled uniformly across all pairs, and as expected, this yields the best results. Since

capturing such a large dataset to calibrate any given camera may be practically burdensome, we also consider subsets derived from a limited number of illuminants, and with a limited number of exposures per-illuminant. The exposures are equally spaced from the lowest to the highest, and the subset of illuminants are chosen so as to maximize the diversity of included chromaticities— specifically, we order the illuminants such that for each n, the convex hull of the RAW R-G chromaticities of patches from the first n illuminants has the largest possible area. The results show that different cameras have different degrees of sensitivity to diversity in exposures and illuminants, but using four illuminants with eight exposures represents a reasonable acquisition burden while also providing enough diversity for reliable calibration in all cameras.

2.5.2 Wide Gamut Imaging

To experimentally compare reconstruction quality of the deterministic and probabilistic approaches in the wide gamut imaging application, we use all RAW-JPEG color-pairs from the database of colors captured with the Panasonic DMC LX-3, corresponding to all color-pairs except those from the four training illuminants. We consider the color checker under a particular illuminant to be the target HDR scene, and we consider the differently-exposed JPEG images under that illuminant to be the input images of this scene. The task is to estimate for each target scene (each illuminant) the true linear patch color from only two differently-exposed JPEG images. The true linear patch color for each illuminant is computed using RAW data from all exposures, and performance is measured using relative RMSE:

$$\text{Error}(\mathbf{x}, \mathbf{x}_{\text{true}}) = \frac{\|\mathbf{x} - \mathbf{x}_{\text{true}}\|}{\|\mathbf{x}_{\text{true}}\|}. \tag{2.17}$$

Figure 2.4 shows a histogram of the reduction in RMSE values when using the probabilistic approach. This is the histogram of differences between evaluating (2.17) with probabilistic and deterministic estimates $\mathbf{x}$ across 1680 distinct linear scene colors in the dataset and all possible un-ordered

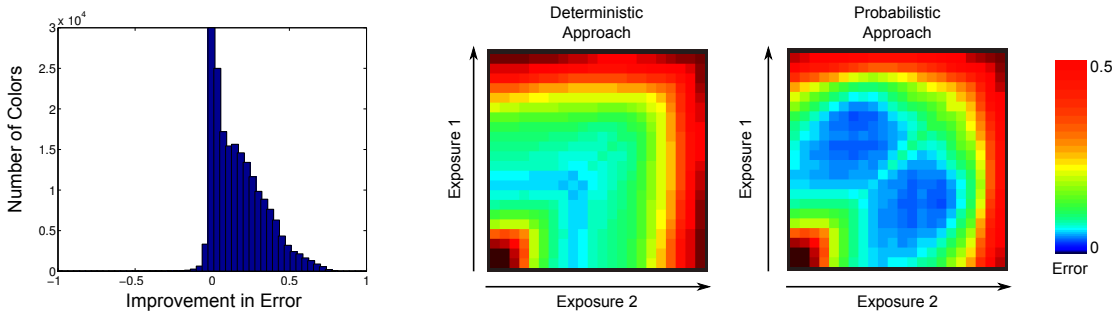


Figure 2.4: Wide gamut imaging results on the Panasonic DMC LX3. (Left) Histogram of improvement in errors over the deterministic baseline for all scene colors using every possible exposure pair. (Right) Mean errors across all colors for each approach when using different exposure pairs.

pairs of 22 exposures[6] as input, excluding the trivial pairs for which $\alpha_1 = \alpha_2$ (a total of 388080 test cases). In a vast majority of cases, incorporating derendering uncertainty leads to better performance.

We also show in the right of the figure, for both the deterministic and probabilistic approaches, two-dimensional visualizations of the error for each exposure-pair. Each point in these visualizations corresponds to a pair of input exposure values (α_1, α_2), and the pseudo-color depicts the mean RMSE across all 1680 linear scene colors in the test dataset. (Diagonal entries correspond to estimates from a single exposure, and are thus identical for the probabilistic and deterministic approaches). We see that the probabilistic approach yields acceptable estimates with low errors for a larger set of exposure-pairs. Moreover, in many cases it leads to lower error than those from either exposure taken individually, demonstrating that the probabilistic modeling is not simply selecting the better exposure, but in fact combining complementary information from both observations.

2.5.3 Photometric Stereo

Finally, we evaluate our model in the context of probabilistic Lambertian photometric stereo. We use JPEG images of a figurine captured using the Canon EOS 40D from a fixed viewpoint under

[6] These correspond to the different exposure time stops available on the camera: $[5e{-}4, 6.25e{-}4, 1e{-}3, 1.25e{-}3, 2e{-}3, 2.5e{-}3, 3.13e{-}3, 5e{-}3, 6.25e{-}3, 1e{-}2, 1.26e{-}2, 1.67e{-}2, 2e{-}2, 2.5e{-}2, 3.33e{-}2, 4e{-}2, 5e{-}2, 6.67e{-}2, 1e{-}1, 2e{-}1, 4e{-}1, 1]$ in relative time units.

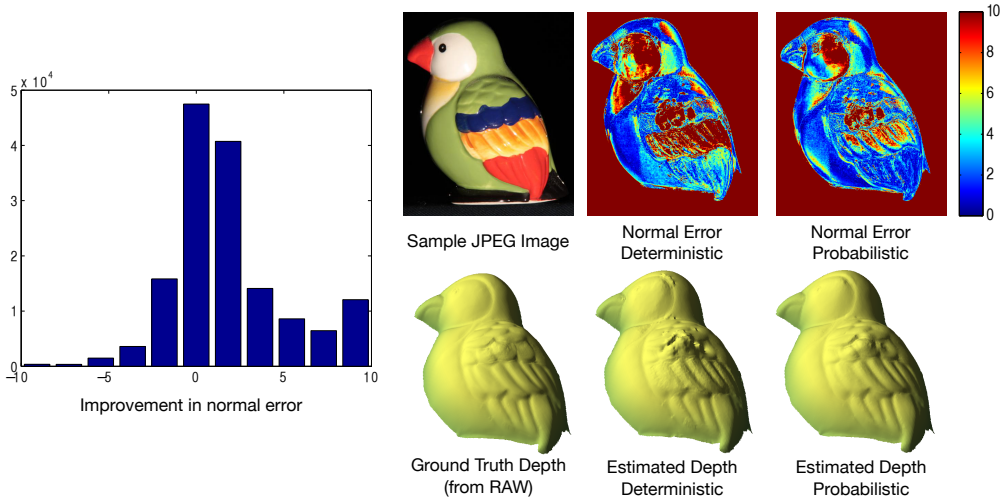


Figure 2.5: Photometric stereo results using the Canon EOS 40D. (Left) Histogram of the improvement in angular error of normal estimate. (Right) One of the JPEG images used during estimation, and angular error (in degrees) for the normals estimated using the deterministic and probabilistic approaches, along with the corresponding depth maps.

directional lighting from ten different known directions. At each pixel, we discard the brightest and darkest measurements to avoid possible specular highlights and shadows, and use the rest to estimate the surface normal. The camera takes RAW images simultaneously, which are used to recover surface normals that we treat as ground truth.

Figure 2.5 shows the angular error map for normal estimates using the proposed method, as well as the deterministic baseline. We also show the corresponding depth maps obtained from the normal estimates using [19]. The proposed probabilistic approach produces smaller normal estimate errors and fewer reconstruction artifacts than the deterministic algorithm—quantitatively, the mean angular error is 4.34° for the probabilistic approach, and 6.46° for the deterministic baseline. We also ran the reconstruction algorithm on inverse estimates computed by simple gamma-correction on the JPEG values (a gamma parameter of 2.2 is assumed). These estimates had a much higher mean error 14.65°.

2.6 Conclusion

Most images captured and shared online are not in linear (RAW) formats, but are instead in narrow-gamut (sRGB) formats with colors that are severely distorted by cameras' color rendering processes. In order for computer vision systems to maximally exploit the color information in these images, they must first undo the color distortions as much as possible. This chapter advocates a probabilistic approach to color de-rendering, one that embraces the multivalued nature of the de-rendering map by providing for each rendered sRGB color a distribution over the latent linear scene colors that could have induced it. An advantage of this approach is that it does not require discarding any image data using ad-hoc thresholds. Instead, it allows making use of all rendered color information by providing for each de-rendered color a measure of its uncertainty.

Our experimental results suggest that a probabilistic representation can be useful when combining per-image estimates of linear scene color, and when recovering the shape of Lambertian surfaces via photometry. The output of our approach—a mean and variance over scene colors for each sRGB image color—may have a practical impact for probabilistic adaptations of other computer vision tasks as well (deblurring, dehazing, matching and stitching, color constancy, image-based modeling, object recognition, *etc*.). One direction worth exploring is the use of spatial structure in the input sRGB image(s), such as edges and textures, to further constrain the de-rendered scene colors. This is in the spirit of [76], and it begs the question of how well a full-gamut linear scene color image can be recovered from a single tone-mapped sRGB one.

3

From Shading to Local Shape

3.1 Introduction

Recovering shape from diffuse shading is point-wise ambiguous because each surface normal can lie anywhere on a cone of directions. Surface normals are uniquely determined only where they align with the light direction which, at best, occurs at only a handful of singular points. A common strategy for reducing the ambiguity is to pursue global reconstructions of large, pre-segmented regions, with the hope that many point-wise ambiguities will collaboratively resolve, or that shape information will successfully propagate from identifiable singular points and occluding contours.

Global strategies are difficult to apply in natural scenes because diffuse shading is typically intermixed with other phenomena such as texture, gloss, shadows, translucency, and mesostructure. Oc-

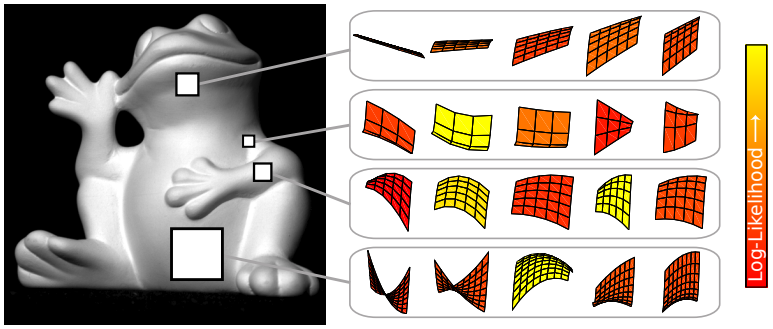


Figure 3.1: We infer from a Lambertian image patch a concise representation for the distribution of quadratic surfaces that are likely to have produced it. These distributions naturally encode different amounts of shape information based on what is locally available in the patch, and can be unimodal (row 2 & 4), multi-modal (row 3), or near-uniform (row 1). This inference is done across multiple scales.

cluding contours and singular points are hard to detect in these scenes; and shading-based shape propagation breaks down unless occlusions, gloss, texture, *etc*. are somehow analyzed and removed by additional visual reasoning. Moreover, most global strategies do not provide spatial uncertainty information to accompany their output reconstructions, and this limits their use in providing feedback to improve top-down scene analysis, or in co-computing with other necessary bottom-up processes that perform complimentary analysis of other phenomena.

This chapter presents theory and algorithms for leveraging diffuse shading more broadly and robustly by developing a richer description of what it says *locally* about shape. We show that point-wise ambiguity can be systematically reduced by jointly analyzing intensities in small image patches, and that some of these patches are inherently more informative than others. Accordingly, we develop an algorithm that produces for any image patch a concise distribution of surface patches that are likely to have created it. We propose these dense, local shape distributions as a new mid-level scene representation that provides useful local shape information without over-committing to any particular image explanation. Finally, we show how these local shape distributions can be combined to recover global object-scale shape.

The main contributions of this chapter are:

1. *Local uniqueness.* We provide uniqueness results for jointly recovering shape and lighting from a small image patch. By considering a world in which the shape of each small surface patch is exactly the graph of a quadratic function in the view-dependent coordinate system, we prove two generic facts: i) when the light direction is known, quadratic shape is uniquely determined; and ii) when the light is unknown, it is determined up to a four-way choice. More interestingly, if the quadratic patch is parametrized intrinsically from the local frame, we show that there are up to four-fold shape ambiguities for any given lighting direction. We also catalog the degenerate cases, which correspond to special shapes, or conspiracies between the light and shape. These results are of direct interest to those studying the mathematics of shape from shading.

2. *Local shape distributions.* We introduce a computational process that takes an image patch at any scale and produces a compact distribution of quadratic shapes that are likely to have produced it. At the core of this process is our observation that all likely shapes corresponding to a (noisy) image patch lie close to a one-dimensional manifold embedded in the five-dimensional space of quadratic shapes. This part of the chapter is of broad interest because these local, multi-scale shape distributions may be useful as intermediate scene descriptors for various visual tasks.

These two parts are tightly bound together. The uniqueness results in Section 3.3 show that the quadratic model is a particularly convenient representation for small surface patches. In the absence of noise, both shape and lighting are locally revealed, local shape is generally unique when lighting is known, and the degenerate cases are easy to describe. Building on this, Section 3.4 examines how uniqueness breaks down in the presence of noise. While very different quadratic shapes can produce equally-likely local intensity patterns, we find that all highly-likely shapes lie close to a one-dimensional sub-manifold. Then, Section 3.5.1 shows how to infer a dense set of sample shapes along this sub-manifold, thereby taking an image patch and producing a one-dimensional shape distribution. The

local model and distribution can potentially be used for many applications. As an example, this chapter also briefly describes one application of these results to the problem of object-scale reconstruction in Section 3.5.2.

3.2 Related Work

Background on shape inference from diffuse shading can be found in several reviews and surveys [15, 33, 94]. An important question is whether shape is uniquely determined by a noiseless image, which has been addressed by a variety of PDE-based formulations. For example, Oliensis considered C^2 surfaces and showed that shape can be uniquely determined for the entire image by singular points and occluding boundaries together [58], and in many parts of the image by singular points alone [57]. For the more general class of C^1 surfaces, Prados and Faugeras [63] employed a smoothness constraint to prove uniqueness properties in a more general perspective setup [62, 64] given appropriate boundary conditions. In this chapter, we use a more restrictive local surface model but prove local uniqueness without any boundary conditions or knowledge of singular points. This generalizes previous studies of local uniqueness, which have considered locally-spherical [61] and fronto-parallel [82] surfaces.

Global uniqueness analyses have inspired global propagation and energy-based methods for global shape inference (*e.g.* [15, 36, 95]), some of which rely on identifying occluding boundaries and/or singular points. While most methods do not typically provide any measurement of uncertainty in their output, progress toward representing shape ambiguity was made by Ecker and Jepson [16], who use a polynomial formulation of global shape from shading to numerically generate distinct global surfaces that are equally close to an input image. In this chapter, we study uniqueness and uncertainty at the local level, and infer distributions over candidate local shapes.

Our work is related to patch-based approaches that use synthetically-generated reference databases. The idea there is to reconstruct depth (or other scene properties [20]) by synthesizing a database of aligned image and depth-map pairs, and then finding and stitching together depth patches from this

database to match the input image and be spatially consistent. Hassner and Basri [28] obtain plausible results this way when the input image and the database are of similar object categories, and Huang et al. [34] pursue a similar goal for textureless objects using a database of rendered Lambertian spheres. Cole et al. [11] focus on patches located at detected keypoints near an object's occlusion boundaries, combining shading and contour cues. We also describe global shape as a mosaic of per-patch depth primitives, but instead of relying on primitives from a pre-chosen set of 3D models, we consider a continuous five-parameter family of depth primitives corresponding to graphs of quadratic functions at multiple scales.

One of the our main motivations is the long-term goal of enabling better co-computation with other bottom-up and top-down visual processes, and by providing useful local shape information without choosing any single image interpretation, our distributions are consistent with Marr's principle of least commitment [53]. We focus on diffuse shading on textureless surfaces, leaving for future work the task of merging with bottom-up processes for other cues like occluding contours (*e.g.*, [11, 35]), texture, gloss, and so on. Our belief that this will be useful is bolstered by promising results achieved by recent global approaches to such combined reasoning [2].

In independent work, Kunsberg and Zucker [42, 43] have recently derived local uniqueness results that are related to, and consistent with, our results in Section 3.3. Their elegant analysis, which uses differential geometry and applies to continuous images, is complimentary to the discrete and algebraic approach employed in this chapter. Kunsberg and Zucker also observe that the analysis of shading in patches instead of at isolated points is consistent with early processing in the visual cortex, and they discuss the possibility of local shading distributions being computed there. Indeed, the notion of such distributions is compatible with evidence that humans perceive shape in some diffuse regions more accurately than others [82].

3.3 Uniqueness Theory in Quadratic-Patch Shape from Shading

We begin by analyzing the ability to uniquely determine the shape and lighting of a local patch from a Lambertian shading image in the absence of noise. The key assumption in our analysis is that depth of the patch can be *exactly* expressed as the graph of a quadratic function. While subsequent sections consider deviations from this idealized setting, the following analysis characterizes the inherent ambiguity under a local quadratic patch model.

We consider two different local coordinate systems for modeling a small surface patch as the graph of a quadratic function: a view-dependent coordinate system where the z axis of the patch is the same as the viewing direction; and a view-independent coordinate system where the z axis is the same as normal vector direction of the local patch. Note that these are two essentially different local models rather than a simple change of parametrization — a patch that can be expressed as the graph of a quadratic function in one coordinate system is not necessarily (in fact, almost never is) the graph of a quadratic function in the other coordinate system. The former local model is more convenient for merging different local patches into a whole surface. We discuss its uniqueness properties in Section 3.3.1 and use it later in the rest of this chapter for instability analysis and surface reconstruction. The latter local model is more natural and intrinsic to the local geometry, but also more mathematically complex. We resort an approximate imaging model that ignores the small surface foreshortening to simplify the analysis, which will be discussed in detail in Section 3.3.2. The results show strong connection to the differential geometry analysis given by Kunsberg and Zucker [42, 43].

3.3.1 Local Quadratic Model in View-Dependent Coordinate System

We model the depth $z(x, y)$ of a local surface patch as a quadratic function defined by coefficient vector $a \in \mathbb{R}^5$ up to a constant offset:[1]

$$z(x, y; a) = a_1 x^2 + a_2 y^2 + a_3 xy + a_4 x + a_5 y. \tag{3.1}$$

In matrix form, this is $z = [x, y]H[x, y]^T + J[x, y]^T$ with

$$H = \begin{bmatrix} a_1 & a_3/2 \\ a_3/2 & a_2 \end{bmatrix} \tag{3.2}$$

the Hessian matrix and $J = [a_4, a_5]$ the Jacobian of the depth function. The un-normalized surface normal to this patch at each location (x, y) is then given by

$$n(x, y; a) = [n_x(x, y; a), n_y(x, y; a), 1]^T, \tag{3.3}$$

where

$$n_x(x, y; a) \triangleq -\frac{\partial z}{\partial x} = -2a_1 x - a_3 y - a_4, \tag{3.4}$$

$$n_y(x, y; a) \triangleq -\frac{\partial z}{\partial y} = -2a_2 y - a_3 x - a_5. \tag{3.5}$$

[1] Local shading for the special case $a_4 = a_5 = 0$ is described in [82], and a more restrictive, locally-spherical model $z(x, y) = \sqrt{r^2 - x^2 - y^2}$ is analyzed in [61].

In matrix form, this is $n(x, y; a) = A[x, y, 1]^T$ with

$$A \triangleq \begin{bmatrix} -2a_1 & -a_3 & -a_4 \\ -a_3 & -2a_2 & -a_5 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.6}$$

the shape matrix corresponding to quadratic shape a.

The intensity $I(x, y; a)$ of this patch, observed from viewing direction $v = [0, 0, 1]^T$ under a directional light source $l = [l_x, l_y, l_z]^T$, is

$$I(x, y; a) = \frac{l^T n(x, y; a)}{\|n(x, y; a)\|}, \tag{3.7}$$

assuming spatially-uniform Lambertian reflectance and that no part of the patch is in shadow, *i.e.*, $l^T n(x, y) > 0, \forall (x, y)$. Here, the magnitude $||l||$ of the light vector represents the product of the surface albedo and the light strength, and it is not assumed to be equal to one. Re-arranging, the intensity I at each point (x, y) induces a quadratic constraint on its surface normal [16]:

$$I^2 n^T n = n^T l l^T n \quad \Rightarrow \quad n^T \left(l l^T - I^2 \mathbb{I}_{3\times 3} \right) n = 0, \tag{3.8}$$

where $\mathbb{I}_{3\times 3}$ is the identity matrix. This further induces a related constraint on shape parameters a:

$$\begin{bmatrix} a^T & 1 \end{bmatrix} \left(D^T \left(l l^T - I^2 \mathbb{I}_{3\times 3} \right) D \right) \begin{bmatrix} a \\ 1 \end{bmatrix} = 0, \tag{3.9}$$

where we use the matrix $D \in \mathbb{R}^{3\times 6}$ to re-write the relationship between n and a in (3.3)-(3.5) as $n = D[a^T \ 1]^T$.

Every pixel (x, y) in an image patch gives one such constraint on shape parameters a, and shape

from shading for quadratic patches rests on solving this system of polynomial equations. Our immediate goal is to determine whether the shape a and lighting l can be uniquely determined from these local constraints.

Uniqueness of simultaneous shape and light

We assume that the local patch is sufficiently large to contain a minimum number of *non-degenerate* pixel locations, where the condition for non-degeneracy is defined as follows:

Definition 3.1. For a patch $\Omega = \{(x_i, y_i)\}_{i=1}^{N}$, we define the matrix $V_\Omega \in \mathbb{R}^{N\times 15}$ such that each row v_i of V_Ω consists of all fourth-order and lower terms of x_i and y_i:

$$v_i = \left[x_i^4,\ x_i^3 y_i,\ \ldots \underset{p,q\geq 0,\ p+q\leq 4}{x_i^p y_i^q} \ldots,\ x_i,\ y_i,\ 1 \right]. \tag{3.10}$$

A patch Ω is considered *non-degenerate* if the matrix V_Ω has rank 15.

Note that rectangular grids of pixels that are 5×5 or larger will be non-degenerate under the definition above.

Theorem 3.2. *Given intensities $I(x, y)$ in an image patch Ω collected at a set of non-degenerate locations not in shadow, if any quadratic-patch/lighting pair (a, l) that satisfies the set of polynomial equations* (3.9) *has a surface Hessian with eigenvalues that are not equal in magnitude, then there are no more than four distinct surfaces that can create the same image. Each of these surfaces is associated with a unique lighting when the Hessian of any solution is non-singular, and a one-dimensional family of lighting vectors otherwise.*

This theorem states that given measurements of intensity from a quadratic surface patch, there generically exists four physical explanations, each comprised of a shape a, a light direction $l/\|l\|$, and a scalar $\|l\|$ encoding the product of albedo and light strength.

Before proceeding to the proof, we introduce a lemma that relates to equations with ratios of quadratic terms. We define $\bar{x} \triangleq [x \;\; y \;\; 1]^T$, so that the normals are given by $n(x, y; a) = A\bar{x}$, and the intensity constraint (3.9) becomes

$$I_{\bar{x}}^2 = \left(\frac{l^T n}{\|n\|}\right)^2 = \frac{\bar{x}^T A^T l l^T A \bar{x}}{\bar{x}^T A^T A \bar{x}}. \tag{3.11}$$

Using this notation, we can state the following lemma, which is proven in Appendix A.1:

Lemma 3.3. *Let A and $\widetilde{A}$ correspond to two matrices of the form in* (3.6)*, and l and $\widetilde{l}$ to two lighting vectors. If*

$$\frac{\bar{x}^T A^T l l^T A \bar{x}}{\bar{x}^T A^T A \bar{x}} = \frac{\bar{x}^T \widetilde{A}^T \widetilde{l}\widetilde{l}^T \widetilde{A} \bar{x}}{\bar{x}^T \widetilde{A}^T \widetilde{A} \bar{x}}, \forall \bar{x} \in \Omega, \tag{3.12}$$

and if Rank$(V_\Omega) = 15$, Rank$(A) \geq 2$, and $l^T A \bar{x} > 0, \forall \bar{x} \in \Omega$ (i.e., no point is in shadow), then

$$A^T l l^T A = \widetilde{A}^T \widetilde{l}\widetilde{l}^T \widetilde{A}, \;\; A^T A = \widetilde{A}^T \widetilde{A}. \tag{3.13}$$

Moreover, if Rank$(A) = 2$, then Rank$(\widetilde{A}) = 2$ and both A and $\widetilde{A}$ share a common null space.

Proof of Theorem 3.2: Suppose there exists a solution (a, l) that produces the observed set of intensities in the patch Ω, and the Hessian matrix of surface a has eigenvalues of un-equal magnitude. We will prove that if there exists another solution $(\widetilde{a}, \widetilde{l})$, such that

$$\frac{\bar{x}^T A^T l l^T A \bar{x}}{\bar{x}^T A^T A \bar{x}} = I_{\bar{x}}^2 = \frac{\bar{x}^T \widetilde{A}^T \widetilde{l}\,\widetilde{l}^T \widetilde{A} \bar{x}}{\bar{x}^T \widetilde{A}^T \widetilde{A} \bar{x}}, \;\; \forall \bar{x} \in \Omega_i, \tag{3.14}$$

then $\tilde{a}$ must be related to a in one of four specific ways.

Since a is not planar (otherwise the Hessian would have both eigenvalues equal to zero), the corre-

sponding matrix A is at least rank 2, and we can apply Lemma 3.3:

$$\widetilde{A}^T\widetilde{l}\,\widetilde{l}^T\widetilde{A} = A^T l l^T A, \qquad \widetilde{A}^T\widetilde{A} = A^T A. \tag{3.15}$$

We define a new matrix B satisfying $\widetilde{A} = BA$. Specifically, when A is full rank, set $B = \widetilde{A}A^{-1}$; and when $\text{Rank}(A) = 2$, set $B = (\widetilde{A}+vv^T)(A+vv^T)^{-1}$ with v a vector in the common null-space of A and $\widetilde{A}$, *i.e.*, $Av = \widetilde{A}v = 0$. We will show that there are only four possibilities for matrix B.

Note that A and $\widetilde{A}$ are affine matrices (last rows are both $[0, 0, 1]$). Moreover, in the rank 2 case, the last entry of v will be 0 and $A + vv^T$ will also be an affine matrix. Therefore, A^{-1} (if A is full rank) and $(A + vv^T)^{-1}$ (if A is rank 2) are affine. Hence, B is also an affine matrix:

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.16}$$

From (3.15), we have $B^T B = \mathbb{I}_{3\times 3}$, *i.e.*,

$$b_{13}^2 + b_{23}^2 + 1 = 1 \quad \Longrightarrow \quad b_{13} = b_{23} = 0. \tag{3.17}$$

The orthogonality of B further restricts its top-left block to be either a 2D rotation matrix

$$B = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.18}$$

or an "anti-rotation" matrix

$$B = \begin{bmatrix} \cos\varphi & \sin\varphi & 0 \\ \sin\varphi & -\cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.19}$$

for $\varphi \in [-\pi, \pi)$.

From $\widetilde{A} = BA$ and the fact that the $(1,2)$-entry and $(2,1)$-entry of $\widetilde{A}$ matrix should be the same (since $a_{12} = a_{21} = -a_3$,$\widetilde{a}_{12} = \widetilde{a}_{21} = -\widetilde{a}_3$), we have

$$2a_1b_{21} + a_3b_{22} = a_3b_{11} + 2a_2b_{12}. \tag{3.20}$$

This implies that when B is of the form in (3.18)

$$(a_1 + a_2)\sin\varphi = 0, \tag{3.21}$$

and when B is of the form in (3.19)

$$(a_1 - a_2)\sin\varphi = a_3\cos\varphi. \tag{3.22}$$

Since the Hessian of a defined in (3.2) has eigenvalues of un-equal magnitude, $a_1 + a_2 \neq 0$, and either

$a_1 \neq a_2$, or $a_3 \neq 0$. This leaves only four possible solutions for B:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos\varphi_0 & \sin\varphi_0 & 0 \\ \sin\varphi_0 & -\cos\varphi_0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} -\cos\varphi_0 & -\sin\varphi_0 & 0 \\ -\sin\varphi_0 & \cos\varphi_0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.23}$$

where $\varphi_0 = \arctan\frac{a_3}{a_1 - a_2}$. Thus $\widetilde{A} = BA$ can relate to A in only four possible ways.

Next, we consider the lighting $\widetilde{l}$ associated with each shape $\widetilde{A}$. Equation (3.15) implies $\widetilde{A}^T\widetilde{l} = A^T l$ or $\widetilde{A}^T\widetilde{l} = -A^T l$ but the latter has shadows, so

$$A^T l = \widetilde{A}^T\widetilde{l} = A^T B^T\widetilde{l}. \tag{3.24}$$

When A is full rank, (3.24) implies a unique $\widetilde{l}$ given by

$$\widetilde{l} = (B^T)^{-1} l = Bl. \tag{3.25}$$

If Rank$(A) = 2$, we define $l_\perp$ as the component of l in the null space of A^T. Then from (3.24),

$$B^T\widetilde{l} = l + cl_\perp \quad \Rightarrow \quad \widetilde{l} = B(l + cl_\perp), \tag{3.26}$$

where c is a scalar. In this case there is a 1D family of $\widetilde{l}$ for each of the four shapes $\widetilde{A}$. □

Figure 3.2 provides an example of the four choices of shape/light pairs in the generic, non-cylindrical

case when both eigenvalues of the surface Hessian are non-zero. Without loss of generality, we consider a rotated co-ordinate system where $a_3 = 0$, *i.e.*, the x and y axes are aligned with the eigenvectors of the surface Hessian. Then, the four solutions from (3.23) are:

$$\big([a_1, a_2, 0, a_4, a_5],\ [l_x, l_y, l_z]\big), \tag{3.27}$$

$$\big([-a_1, -a_2, 0, -a_4, -a_5],\ [-l_x, -l_y, l_z]\big), \tag{3.28}$$

$$\big([a_1, -a_2, 0, a_4, -a_5],\ [l_x, -l_y, l_z]\big), \tag{3.29}$$

$$\big([-a_1, a_2, 0, -a_4, a_5],\ [-l_x, l_y, l_z]\big). \tag{3.30}$$

The first choice is the surface/lighting pair (a, l) that actually induced the image. The second corresponds to the well-known convex-concave ambiguity [61], and is obtained by reflecting both the light and the normals across the view direction. The last two choices (3.29)-(3.30) correspond to performing the reflection separately along each of the eigenvector directions of the Hessian matrix. These form a second concave-convex pair.

When one of the Hessian eigenvalues is zero (say $a_2 = 0$ in our rotated co-ordinate system), the patch surface is a cylinder and it is possible to construct a 1D family of lights for each of the four surfaces:

$$\widetilde{l} = \text{diag}\{\text{sign}(\widetilde{a}_1 a_1), \text{sign}(\widetilde{a}_5 a_5), 1\}\ (l + c \cdot [0, 1, a_5]^T) \tag{3.31}$$

for any $c \in \mathbb{R}$ such that no pixel is in shadow. Figure 3.3 shows an example of four cylindrical surfaces and associated families of lights that can produce the same image.

Theorem 3.2 applies when the Hessian eigenvalues of any solution shape are not equal in magnitude. What happens when shape solutions have Hessian eigenvalues that *are* of equal magnitude? There are two distinct cases. The first is when the Hessian is zero and the true surface is planar. In this case every surface normal in the patch is identical, and the well-known point-wise cone ambiguity

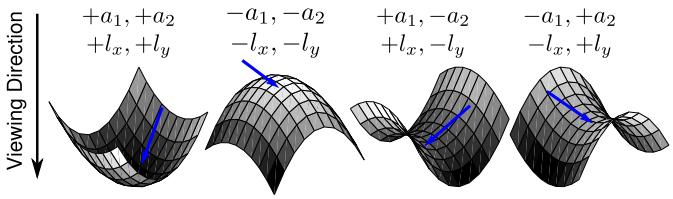


Figure 3.2: Four quadratic-patch/lighting configurations that produce the same image (left is $a = [1, 1/2, 0, 0, 0], l = [2/3, 1/3, 2/3]$). The lighting is shown as blue arrows. The left pair and right pair are each convex-concave.

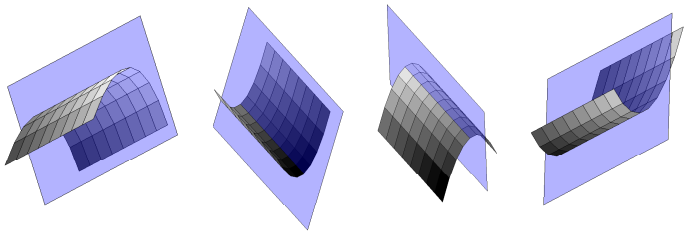


Figure 3.3: Lighting solutions in the cylinder case, when one of the eigenvalues of the surface Hessian is zero. There is a 1D family of lighting (any lighting direction in the blue plane with appropriate strength) for each of the four shapes that can produce the same image.

applies to the patch as a whole: The observed image can be explained by a one-parameter family of planar surfaces for *every* light l.

In the second case, the true surface is not planar but the magnitudes of the two eigenvalues of the Hessian matrix are equal. Unlike the planar ambiguity, there is not an infinite number of surfaces that can combine with every lighting. But as depicted in Figure 3.4, there is still an infinite number of allowable patch/lighting pairs. We note that all quadratic surfaces in this category can be expressed as either one of two following forms

$$a = [r\cos\theta, -r\cos\theta, 2r\sin\theta, p\cos\theta - q\sin\theta, p\sin\theta + q\cos\theta], \tag{3.32}$$

$$a = [\lambda r, \lambda r, 0, \lambda p, -\lambda q], \tag{3.33}$$

where $\theta \in (-\pi, \pi], \lambda \in \{-1, +1\}, r \in \mathbb{R}^+$, and $p, q \in \mathbb{R}$. Given fixed values of r, p and q, these

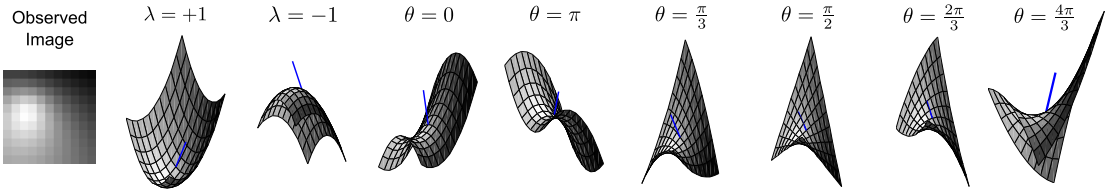


Figure 3.4: When Hessian eigenvalues are equal in magnitude, there is a continuous family of patch/lighting pairs (given by (3.32) and (3.33)) that produce the same image. Note that the first four pairs above are analogous to Figure 3.2.

surfaces generate identical images when paired with lighting

$$l = [l_x \cos\theta - l_y \sin\theta, l_x \sin\theta + l_y \cos\theta, l_z], \tag{3.34}$$

for surfaces (3.32), or with

$$l = [\lambda l_x, -\lambda l_y, l_z], \tag{3.35}$$

for surfaces (3.33), with fixed values of l_x, l_y, l_z.

Unique shape when light is known

Theorem 3.4. *Given intensities $I(x, y)$ at a non-degenerate set of locations Ω, a known light l, and a quadratic patch a that satisfies the set of equations in* (3.9)*, if the planar component $[l_x, l_y]$ of the light is non-zero (i.e, l is not equal to the viewing direction) and not an eigenvector of the Hessian of a, then the solution a is unique.*

Proof of Theorem 3.4: Without loss of generality, we choose a co-ordinate system where $a_3 = 0$. Note that for any such choice l_x and l_y will both be non-zero, unless $[l_x, l_y]$ is zero or an eigenvector of the surface Hessian, which is ruled out by the statement of the theorem.

If the Hessian of a has eigenvalues with unequal magnitudes, then it is easy to see that each of the four possible solutions from Theorem 3.2 has distinct light from (3.25) and (3.26), and therefore for a fixed light, the shape is unique. A Hessian with *equal* eigenvalues is ruled out since then every light-

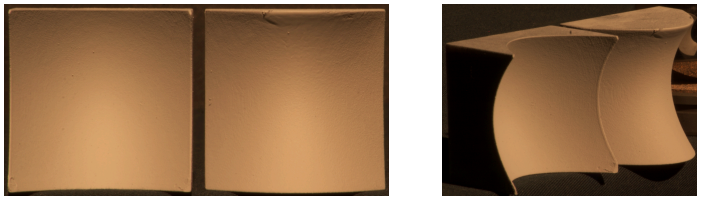


Figure 3.5: Left: Two quadratic surfaces that produce the same image when the light is aligned with one of their common Hessian eigenvectors. For other view and light configurations (*e.g.*, right) their images are distinct.

direction would be an eigenvector. When the eigenvalues have equal magnitudes but opposite signs, a must be of the form in (3.32) with $\theta = 0$ or π (since $a_3 = 0$) and $r = |a_1| = |a_2|$. In this case too, we see that each member of the continuous family of solutions—with $\theta \in (-\pi, \pi]$ for surface (3.32) and light (3.34), or $\lambda \in \{-1, +1\}$ for surface (3.33) and light (3.35)—has a distinct light-direction. □

When the conditions in Theorem 3.4 are not satisfied, there are shape ambiguities as follows. First, planar patches have Hessians with zero eigenvalues so that every l is an eigenvector; this leads to an infinite set of planar shape explanations for any given light. Second, when the light and view directions are the same, there are generically four shape solutions analogous to Figure 3.2 or, in the case of equal eigenvalue magnitudes, a continuous family of solutions analogous to Figure 3.4. Finally, when the true surface is not planar but the azimuthal component of the light $[l_x, l_y]$ happens to be aligned with one of the Hessian eigenvectors, it is possible to construct a second solution by performing a reflection of the normals across that eigenvector direction. Figure 3.5 demonstrates this with photographs of two 3D-printed surfaces that are distinct but related by a horizontal reflection of their normals.

3.3.2 Local Quadratic Model in Intrinsic Coordinate System

In this section we adopt a more intrinsic coordinate system for local patch model. Define the z axis to be the normal vector direction of the patch, making it independent to the external viewing direction. The definition of x and y axes is still dependent on the lighting direction, but that dependency is for representational convenience only — one can arbitrarily rotate these local horizontal axes without

changing the conclusion of analysis in this section.

More specifically, given a lighting source l, for any generic[2] second order smooth patch with normal vector $\mathbf{n}_0 \neq l$, define a local frame $\boldsymbol{F} = [\mathbf{f}_s, \mathbf{f}_t, \mathbf{n}_0]$ such that l lies on the plane spanned by $\mathbf{f}_s$ and $\mathbf{n}_0$. We use (s, t, r) index a point in this local frame, *i.e.*

$$\mathbf{x} = s\mathbf{f}_s + t\mathbf{f}_t + r\mathbf{n}_0 = [\mathbf{f}_s, \mathbf{f}_t, \mathbf{n}_0] \begin{bmatrix} s \\ t \\ r \end{bmatrix} = \boldsymbol{F} \begin{bmatrix} s \\ t \\ r \end{bmatrix}. \tag{3.36}$$

We call a surface patch *quadratic in the intrinsic frame* (or *intrinsic quadratic* in short) if the patch $\mathcal{P}$ can be parametrized by two parameters s and t such that

$$\mathcal{P}(s,t) = s\mathbf{f}_s + t\mathbf{f}_t + \left(a_1 s^2 + a_2 t^2 + a_3 st\right)\mathbf{n}_0 = \boldsymbol{F} \begin{bmatrix} s \\ t \\ a_1 s^2 + a_2 t^2 + a_3 st \end{bmatrix}. \tag{3.37}$$

An intrinsic quadratic patch has five parameters, two (θ, ϕ) parametrize the normal direction $\mathbf{n}_0$ (and thereby the local frame $\boldsymbol{F}$), and the other three (a_1, a_2, a_3) parametrize the curvature of the patch. When we have knowledge of lighting l and an exact measurement of the center pixel intensity, one normal parameter ϕ is locked, and there are four remaining parameters, denoted as $\mathcal{P}(s, t; \theta, a_1, a_2, a_3)$, or $\mathcal{P}(\theta, a_1, a_2, a_3)$ in short when there is no confusion.

[2] There are several generic assumptions made in this section. When we say "generic", it indicates such condition will apply to all surfaces or lights *except* for a measure-zero subset. We will be explicitly state these assumptions throughout the section when they are used.

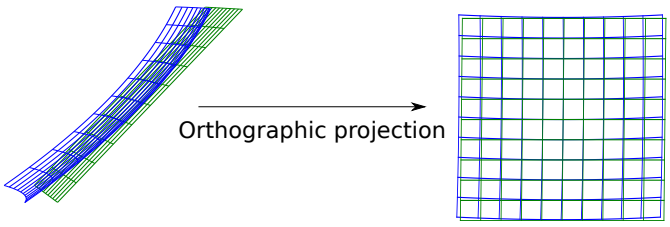


Figure 3.6: A visualization of the approximate imaging model in Definition 3.5. The model takes the intensity values on the curved blue grid and ``glue'' them onto the planar green grid. When orthographically projected onto the image plane, we get the linearized affine grid (green) instead of the non-linear curved grid (blue), which significantly simplifies the calculation.

The (un-normalized) normal vectors for a intrinsic quadratic patch can be written as

$$\mathbf{n}(s,t) = \begin{bmatrix} \mathbf{f}_s & \mathbf{f}_t & \mathbf{n}_0 \end{bmatrix} \begin{bmatrix} n_s \\ n_t \\ 1 \end{bmatrix} = \boldsymbol{F} \begin{bmatrix} -2a_1 s - a_3 t \\ -a_3 s - 2a_2 t \\ 1 \end{bmatrix}. \tag{3.38}$$

When lit with directional lighting l, a Lambertian intensity value can be calculated at each point on $\mathcal{P}(s,t)$ as

$$I(s,t) = \frac{n_s l_s + l_n}{\sqrt{n_s^2 + n_t^2 + 1}}, \tag{3.39}$$

where $l_s = l \cdot \mathbf{f}_s$ and $l_n = l \cdot \mathbf{n}_0$ (note that by definition, $l \cdot \mathbf{f}_t = 0$).

This intrinsic coordinate system, however, is mathematically complex, because the intensity $I(s,t)$ is nonlinearly attached to a surface point $\boldsymbol{F}[s,t,a_1 s^2 + a_2 t^2 + a_3 st]^T$, and there is no direct mapping to the actually observed pixel on the image plane $I(x,y)$. To simplify it, we resort to an approximate imaging model defined below, which "hallucinates" a linear mapping from the local frame (s,t) to the image plane (x,y).

Definition 3.5. When the surface $\mathcal{P}(s,t)$ is lit with directional lighting l, we project the Lambertian intensity value $I(s,t)$ defined in (3.39) from the surface point $\boldsymbol{F}[s,t,a_1 s^2 + a_2 t^2 + a_3 st]^T$ to the tangent surface of patch $\boldsymbol{F}[s,t,0]^T$, resulting in a linear relationship to a pixel in image plane

$I(x, y) = I(\boldsymbol{P}^{-1}(s, t))$, where the 2×2 matrix $\boldsymbol{P}$ is the upper-left block of $\boldsymbol{F}$. We refer to this as the approximate imaging model for intrinsic quadratic patches.

With this imaging model, we present our main uniqueness results for local quadratic model on intrinsic coordinate system below. The detailed proof is given in Appendix A.2, but we make an important remark on the connection of this theorem to the results of previous section here.

Theorem 3.6. *Given intensities $I(x, y)$ in an image patch Ω, generated by approximate imaging model in Definition 3.5 from a patch/lighting pair $(\mathcal{P}(\theta, a_1, a_2, a_3), l)$, then generically for any given lighting $\widetilde{l}$, there are at most four intrinsic quadratic patches $\mathcal{P}(\widetilde{\theta}, \widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3)$ that can exactly explain the image.*

Remark 3.7. Theorem 3.6 is qualitatively different from Theorem 3.2 in Section 3.3.1. Theorem 3.2 states that given a shading patch generated by an extrinsic quadratic surface, there are up to four-fold ambiguity without knowledge of lighting direction, whereas Theorem 3.6 states that the four-fold ambiguity is based on a *given lighting*, and implies a 2D continuous ambiguity without knowledge of lighting. This qualtitative difference in results can be attributed two major modeling difference summarized below:

1. The surface geometric models are different. As noted at the beginning of Section 3.3, the change of coordinate system in local models is not a trivial re-parametrization: the spaces of quadratic patches in the extrinsic and intrinsic coordinate systems are essentially different. Patches that is quadratic in the extrinsic coordinate system rarely live in the quadratic patch space of the intrinsic coordinate system. The local shading images created by patches from either space can therefore differ significantly.

2. The imaging models are also different. The imaging model for Section 3.3.1 is standard Lambertian without any additional assumption or approximation; whereas the imaging model in this section introduces an approximation in Definition 3.5 (see also Figure 3.6). This approxi-

mation ignores the second-order foreshortening effect on the patch curvature, which could be a weak but essential signal that leads to stronger uniqueness results if not ignored.

3.3.3 Connection to Uniqueness Results in Related Work

Uniqueness analysis on surface shape from local shading dates back to the classic work of Pentland [61], where a more restrictive local model that local patches have equal principal curvatures is assumed. In the view-dependent coordinate system, the depth map of this restrictive model can be written as part of sphere of radius r: $z(x, y) = \sqrt{r^2 - x^2 - y^2}$. We quote the main results proved by Pentland in [61] as follows: "if the principal curvatures [of the local surface] are assumed to be equal, there is a unique combination of image formation parameters (up to a mirror reversal) that will produce a particular set of image intensity and first and second derivatives". Here the "image formation parameters" refers to the shape parameter r and the lighting parameters l. Although the local model and proved uniqueness properties are different from ours (2-fold ambiguity rather than 4-fold), we actually share the same essence: both results imply that given the local intensities and restricting the local surface into a strict parametric family, the shape and lighting will be strongly constrained, or uniquely constrained up to a finite (2 or 4) fold of ambiguity.

The four fold ambiguity shown in Theorem 3.2 is in fact a generalization of the classic convex-concave ambiguity in [61]. This has been previously documented in the work by Wagemans et al. [82], who constructs and illustrates the frontal-parallel version of such phenomenon for perception study (see Figure 3 of [82]). This can be thought as a special case of our results, where the surface patch has zero first order derivative and can be written as $z(x, y) = a_1x^2 + a_2y^2 + a_3xy$, equivalent to set $a_4 = a_5 = 0$ in (3.1).

In more recent independent work, Kunsberg and Zucker [42, 43] derived local uniqueness results from differential geometry analysis that are related to and consistent with ours under the intrinsic local coordinate system. The derivation is from a continuous perspective, and we summarize their

main results (Theorem 4.1 in [42]) as following: for any point in the image plane p, the second order derivative of image intensity I_{uu}, I_{uv}, I_{vv} (u, v are the unit length vectors on the image plane, with u aligned to the direction of brightness gradient and v in the direction of isophote and therefore perpendicular to brightness gradient) cast three constraints on the normal vector and curvature of the surface geometry, and these constraints are invariant to the lighting direction. They further argue that for a quadratic patch with 5 degrees of freedom, these 3 constraints will leave a 2D continuous ambiguity. On the other hand, assuming prior knowledge of the normal direction, we will be left with only 3 curvature unknowns, and the constraints on these 3 curvature unknowns become a 4-th order polynomial, which can be solved up to a four-fold ambiguity (see Section 5.3 and Section 5.4 of [42] for more details). These results share strong connection to Theorem 3.6 of previous section. We list the similarities and differences of two approaches below:

The local patch models are essentially the same. In Theorem 3.6, we use an intrinsic local coordinate system of the patch, and explicitly state that the patch can be expressed as the graph of quadratic function $t = a_1 r^2 + a_2 s^2 + a_3 rs$, whereas Kunsberg and Zucker uses a differential geometry setup and looks into the major and minor curvature of the patch, which is exactly the eigenvalues of Hessian matrix $H = [a_1, a_3/2; a_3/2, a_2]$ in our setup.

The imaging model and input intensity representation of the algorithms are different. Our method assumes an approximate imaging model that ignores curvature foreshortening (see Definition 3.5) and uses the intensities at a set of discrete pixels (say a 5×5 grid) as input; whereas Kunsberg and Zucker employ a continuous setup and use the second order intensity derivatives as input, and they do not ignore any foreshortening effect in the analysis. These two sets of input are different because the curvature foreshortening approximation in our imaging model (ignoring quadratic term $a_1 s^2 + a_2 t^2 + a_3 st$ in Equation (3.37)) will change the second order intensity derivatives at center pixel. It can be shown that for a given lighting, the four quadratic surface solutions predicted by our theory that create the same intensities on the discrete grid when ignoring curvature foreshortening will *n*ot have the same

second order intensity derivatives if the curvature foreshortening is not ignored. It remains an interesting future work direction to explore whether our uniqueness theory will still hold for intrinsic quadratic patches without any foreshortening approximation in the imaging model.

The prior knowledge assumed by the two approaches are different and complementary to each other. Our approach considers known lighting direction and shows uniqueness of the entire surface parameters (first and second order), whereas Kunsberg and Zucker assumes the surface normal *a priori*, and reasons about the uniqueness of second order surface parameters that are invariant to lighting directions. As a speculation, we believe that given additional intensity and first order derivatives information to the latter framework, the lighting direction can also be locked down up to a four-fold ambiguity. If that is indeed the case, the results we get are "dual" to Kunsberg and Zucker, in the sense that the dot product of normal and lighting produces the intensity of center pixel, given which knowledge of one vector will confine the other vector to a one-dimensional cone. The uniqueness analysis further reduce the vector on each cone to a finite number of (up to 4) possibilities given additional intensity information of the local neighborhood.

3.4 Ambiguity in the Presence of Noise

The uniqueness results from the previous section suggest that among the many possible models one could use for local shapes—such as splines, linear subspaces, exemplar dictionaries [34], or continuous functions with smoothness constraints as in [2]—the quadratic function model may be particularly useful. However, before we can use this model for inference, we must understand the effects of deviations, such as intensity noise and higher-order (non-quadratic) components of local shape. To this end, we provide some intuition about the types of quadratic shapes that *almost* satisfy the polynomial system (3.9) and thus become likely explanations in the presence of noise. These intuitions motivate a statistical inference technique that will be introduced in Section 3.5.

In the rest of this chapter, we assume that the light direction $l/\|l\|$ and the albedo/light-strength

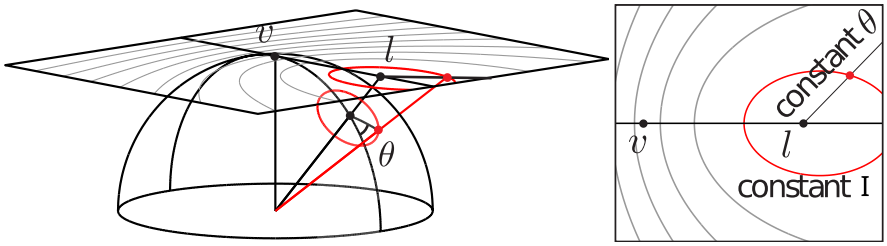


Figure 3.7: The light-centered cone of possible surface normals at any image point projects radially to a conic on the projective plane. We parameterize these conics by the radial projection of spherical angle θ.

product $\|l\|$ are known. Then, the polynomial system (3.9) relating the quadratic parameters a to the observed intensities I can be understood as combining two types of constraints on the patch normals $n = [n_x, n_y, 1]$. First, each pixel's normal is constrained by its intensity to a light-centered circle of directions as per (3.7). This is shown in the left of Figure 3.7, where the circle of directions is parameterized by "azimuthal" angle

$$\theta = \arctan\left(\frac{n_x l_y - n_y l_x}{l_x^2 + l_y^2 - l_z\left(n_x l_x + n_y l_y\right)}\right). \tag{3.40}$$

The second type of constraint comes from the quadratic shape model, which induces a joint geometric constraint on the set of surface normals that belong to the patch. This joint constraint has an intuitive interpretation when we represent the normals, light, and view as points on the plane defined by $n_z = 1$ (the so-called projective plane [79]). This representation is constructed by radially-projecting the hemisphere of directions onto the plane as shown in Figure 3.7. The view is the origin of the plane, the light projects to another planar point $(l_x, l_y)/l_z$, and each pixel's θ-parameterized circle of normal azimuthal directions projects to a conic section, still parameterized by θ. The set of normals that lie on different conics but have the same azimuthal angle θ form a ray (right of Figure 3.7), and an inversion in the sign of θ corresponds to a reflection of the surface normal across light point.

Using this representation, Figure 3.8 visualizes the two types of constraints (under a light with $l_y = 0$) for 25 normals at a 5×5 grid of (x, y) pixel locations. In addition to each pixel's normal

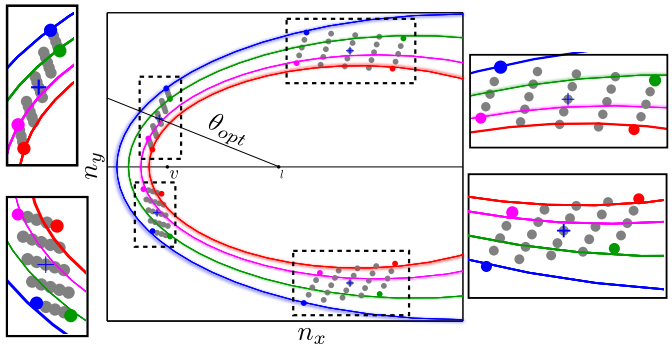


Figure 3.8: Exact and approximate solutions for quadratic shape. Each color corresponds to a pixel in the patch (four are shown in the plot), whose intensity defines a conic curve that the normal vector should lie on. The normal vectors for a quadratic patch should form an affine grid on the projective plane, and good-fit shapes have grids that are well-aligned with the corresponding conics. The top left grid corresponds to an exact fit.

being constrained to its conic, the set of normals is collectively constrained, via (3.6), to be a symmetric affine grid. Therefore, solving the polynomial system for quadratic coefficients a amounts to finding a symmetric affine grid that aligns properly with the per-pixel conics. Theorem 3.4 tells us there is only one grid that aligns perfectly, but as shown in the figure, there will be other grids that come close. When there is noise, the shapes corresponding to all of these grids become likely explanations, even though they are physically quite different from one another. To avoid over-committing, local inference systems must output distributions of shapes that encode this fact.

Then, a natural question is: do we need to search the entire five-dimensional space of quadratic parameters a to find all the likely approximate solutions? To answer this question, we note that these approximate solutions are intuitively expected to arise from the degenerate cases detailed in Theorem 3.4. For example, we find that these solutions often occur in pairs corresponding to reflections across the light direction (*i.e.*, across the x axis in Figure 3.8), which would correspond to a second exact solution if the light were a eigenvector of the surface Hessian. Remember that the most ambiguous degeneracy is the one induced by the true surface being planar, when all the conics overlap and there is a continuous set of solutions whose normals can be parametrized by a single angle θ as per (3.40). Based on this intuition, we define $\theta(a)$ as the first-order orientation of the shape a to be the angle

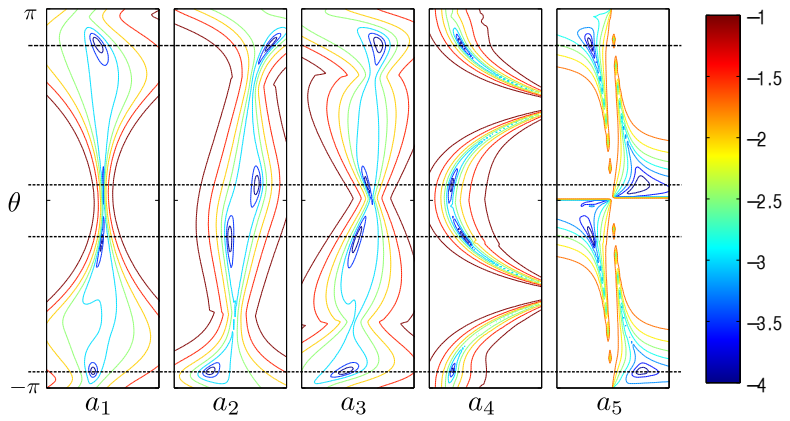


Figure 3.9: Iso-contours of RMS intensity error for renderings of best-fit shape parameters $(a_1, a_2, a_3, a_4, a_5)$ when θ is fixed. Close fits occur at very different orientations (four modes here), but for any fixed orientation θ the remaining shape parameters are very constrained.

of the center normal, and find empirically that it is sufficient to search along only a one-dimensional manifold parametrized by this angle.

In Figure 3.8, this search can be understood as fixing the value of $\theta(a)$, and warping an affine grid by optimizing the parameters a_1, a_2, a_3, a_4, a_5 to fit the conic intensity constraints. We see that this leaves very little play in the parameters, so the shapes a of possible solutions are highly constrained once $\theta(a)$ is fixed. This effect is further visualized in Figure 3.9, which shows contours of constant RMS intensity difference—equally spaced in value on a logarithmic scale—between the observed intensities and the Lambertian renderings of best-fit shapes obtained by fixing $\theta(a)$ and one coefficient (say, a_1) and then optimally fitting the others (say, a_2, a_3, a_4, a_5). The four "close fits" appear as the four modes, where the value of $\theta(a)$ strongly constrains each coefficient of low-error shapes a.

An interesting observation about Figure 3.8 and Figure 3.9 is that there are four distinct local minimum in the one-dimension sub-manifold, and this in fact is not a coincidence. In practice, we found that for most quadratic patches, there are usually two or four strong modes in the one-dimensional θ sub-manifold, which we believe can be attributed to the four-fold ambiguity of the intrinsic quadratic patch proved in Theorem 3.6, although the space of patches we are studying is extrinsically quadratic.

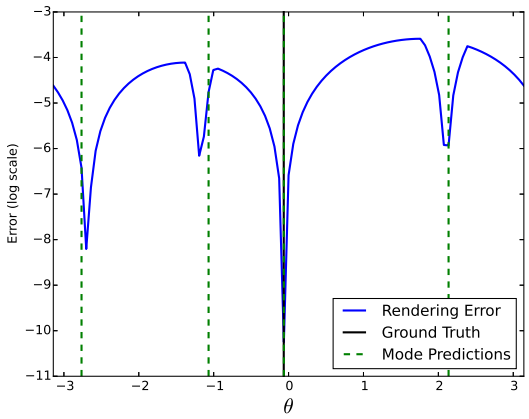


Figure 3.10: Mode prediction with intrinsic quadratic model. The blue curve shows the RMS intensity error for renderings of best-fit extrinsic quadratic model, and the green dotted lines show the mode predictions from intrinsic quadratic model described in Section 3.3.2. The black vertical line is the ground truth θ that generate the input image.

More specifically, given the intensities generated from an extrinsic quadratic patch a and known lighting l, according to Theorem3.4, generically, the only extrinsic quadratic patch that can exactly explain these intensities is a itself. However, in presence of noise, the intrinsic approximation of patch a (referred to as a') also comes close to explain the image intensities (note again that the extrinsic and intrinsic quadratic patches form two different families). Under Theorem3.6, there are up to three other intrinsic quadratic patches that create the exact same intensities to that of a' under lighting l, and therefore these intrinsic patches can also closely explain the input image intensities. Finally, the extrinsic approximation of these three intrinsic patches will produce very similar image as the input.

To verify this intuition, we did the following experiment: create an intensity patch from a given surface patch a, and densely sample the one-dimensional θ sub-manifold to find the best extrinsic quadratic patches that best explain the given intensity. Then convert the ground truth patch a into the intrinsic coordinate system as a' (with some loss of accuracy), find the three other intrinsic patches that generate the same intensities as a' (refer to Appendix A.2 for how the exact calculation is carried out), and convert these patches into extrinsic quadratic space. We examine the θ of the converted-

back patches and see whether they correlate to the best samples in the one-dimensional extrinsic submanifold. The result is shown in Figure 3.10, which affirms our intuition.

3.5 Local Shape Proposals and Surface Reconstruction

Armed with intuition about the characteristics of approximate solutions for the quadratic-patch model, we now develop a method for inferring shape distributions at any local image patch of any size. The output for each image patch is a set of quadratic shapes of the same size that correspond to a discrete sampling along a θ-parametrized one-dimensional manifold, as well as a probability distribution over this set of quadratic shapes. The previous sections have demonstrated that shading in some image patches is inherently more informative than others. Our goal is to create a compact description of this ambiguity in each local region at multiple scales, thereby providing a useful mid-level representation of "intrinsic" scene information for vision.

3.5.1 Computing Quadratic Shape Proposals

Given the intensities $I_o(x, y)$ at a patch $(x, y) \in \Omega$, we first generate a set of quadratic proposals for the shape of that patch, and based on the intuition from the previous section, we index these proposals angularly in reference to the light l. Consider a discrete set of uniformly-spaced values $\theta^j,\ j \in \{1, \ldots J\}$ over $(-\pi, \pi]$[3], and for each angle θ^j we find the corresponding quadratic shape a^j that best explains the observed intensities $I_o(x, y)$ in terms of minimum sum of squared errors:

$$a^j = \underset{a:\theta(a)=\theta^j}{\arg\min} \sum_{(x,y)\in\Omega} \|I_o(x, y) - I(x, y; a)\|^2, \tag{3.41}$$

where $I(x, y; a)$ is defined as per (3.7).

[3]For some patches, we consider closer-spaced samples over a shorter interval when values close to $\pm\pi$ do not correspond to physically feasible estimates for shape.

Let $(0, 0)$ be the center of the patch. Then since $\theta(a_i)$ is fixed, the quadratic coefficients a_4 and a_5 of a^j only have one degree of freedom, and can be re-parametrized in terms of a single variable $r \in \mathbb{R}^+$ that indexes points along the constant θ ray on the projective plane:

$$a_4 = -\frac{l_x}{l_z} - r\left(-\frac{l_x}{l_z}\cos\theta^j + l_y \sin\theta^j\right), \tag{3.42}$$

$$a_5 = -\frac{l_y}{l_z} - r\left(-\frac{l_y}{l_z}\cos\theta^j - l_x \sin\theta^j\right). \tag{3.43}$$

Therefore, the non-linear least-squares minimization in (3.41) is over the four variables $a_{1:3}$, r, and can be efficiently carried out with Levenberg-Marquardt [52]. We found empirically that it is insensitive to initialization, and use $[0, 0, 0, r_0]$ in our experiments, where r_0 is chosen such that the center pixel lies on the corresponding conic.

This minimization occurs independently and in parallel for every patch in an image, and it can therefore be parallelized over an arbitrary number of CPU cores, on a single machine or a cluster of machines, as required for increasing image sizes. Our reference implementation considers $J = 21$ quantized angles for each patch, and takes one minute on an eight-core machine for inference on all overlapping 5×5 patches in a 128×128 image.

3.5.2 Surface Reconstruction

In collaboration with Ayan Chakrabarti, we demonstrate the utility of our theory and local distributions for higher-level scene analysis by reconstructing object-scale surface shape when the light l is known. The local representations provide concise summaries of the shape information available in each image patch, and they do this without "over-committing" to any one local explanation. This allows us to achieve reliable performance with very a simple algorithm for global reasoning that infers object-scale shape through simple iterations between: 1) choosing one likely shape proposal for each local patch; and 2) fitting a global smooth surface to the set of chosen per-patch proposals.

Figure 3.11: Surface reconstruction on real captured data. We show two novel view points for each reconstruction, and the median angular error between estimated surface normal vectors and ground truth surface normal vectors.

Figure 3.11 shows the surface reconstruction results of the proposed algorithm, in comparison with two state-of-the-art methods. The first is the iterative algorithm proposed by Ecker and Jepson [16] (labeled "Polynomial SFS"). The second (labeled "Cross-scale") is the shape from shading component of the SIRFS method [2], *i.e.*, where we treat the light and shading-image as given, and do not use contour information. More details of the proposed algorithms can be found in our journal paper [86].

3.6 Discussion

Our theoretical analysis shows that in an idealized quadratic world, local shape can be recovered uniquely in almost every local image patch, without the use of singular points, occluding contours, or any other external shape information. Beyond this idealized world, our evaluations on synthetic and captured images suggest that one can infer, efficiently and in parallel, concise multi-scale local shape distributions that are accurate and useful for global reasoning.

There are many viable directions for interesting future work. Foremost among these is the joint estimation of shape, lighting, and albedo. The reconstruction algorithms proposed in this chapter are limited to the case when lighting is known, but the uniqueness results in Section 3.3.1 suggest that simultaneous reconstruction of shape and lighting may also be possible. Theorem 3.2 tells us that, in an idealized quadratic world, there are generically four lights l that can explain each local patch, and that these quadruples of possible lights will vary from patch to patch according to the directions of each patch's Hessian eigenvectors. Intuitively, one might infer the true light (along with its reflection across the view, which is always equally-likely) as the one that is common to all or most of the per-patch quadruples.[4] Practically speaking, it is likely that for a reconstruction algorithm to handle unknown lighting, it will need to jointly reason about shape, lighting, and varying albedo, in the same

[4] We have experimented with a direct implementation of this intuition that does a brute-force search only on lighting direction, assuming a known constant light-strength and albedo, and with pooling local estimates without considering consistency or noise. This method worked reasonably well in many cases, but was computationally expensive and not entirely robust.

spirit as Barron and Malik [2]; and that such reasoning will benefit from an analysis of the joint ambiguities that are induced by noise and non-quadratic shape, similar to what was done for shape alone in Sections 3.4 and 3.5.

Also, while we provide a means to extract a single estimate of the global surface from local shape distributions, one could also imagine using reasoning about consistency and outliers to allow the full distributions of neighboring patches to collaboratively refine themselves. This could be useful, for instance, when the object boundaries in a scene are not known a-priori. These refined local distributions may then be able to identify depth discontinuities in the scene, and help segment out individual objects for shape recovery.

Finally, it will be interesting to pursue combining our shading-based local distributions with complementary reasoning about contours, shading keypoints [26], texture, gloss, shadows, and so on—treating these as additional cues for shape, as well as to better identify outliers to our smooth diffuse shading model. We also believe it is worth integrating these local shape distributions into processes for higher-level vision tasks such as pose estimation, object recognition, and multi-view reconstruction, where one can imagine additionally using top-down processing to aid local inference, for example by exploiting priors on local quadratic shapes that are based on object identity or scene category.

4

Consensus of Regions in Spatial Hierarchy

4.1 Introduction

Physics-based visual inference, also known as low-level vision, is the estimation of depth, motion, shape, and other physical scene properties from visual measurements. Since it is ill-posed, methods often employ a *local model* that is expected to apply piecewise across the scene, and that restricts the variation of scene values within each applicable piece or region. Slanted planes for binocular disparity, constant or affine optical flows, and families of smooth shapes for surface normals are common examples (see Figure 4.1). The restriction on scene variability in applicable regions allows image cues to be aggregated spatially across each region, thereby reducing the ambiguity that exists point-wise. The fundamental challenge lies in identifying—automatically from the image input—the sizes and

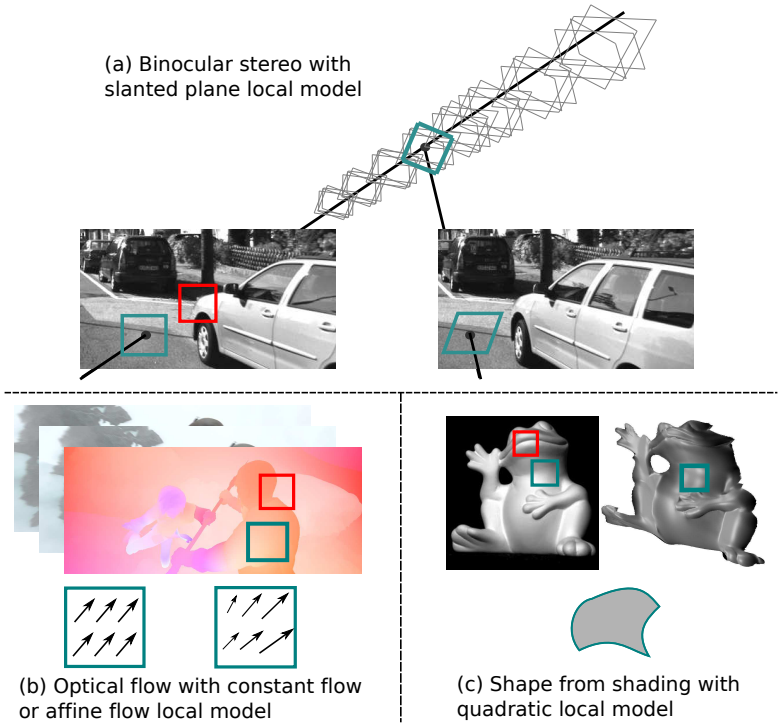


Figure 4.1: Local models for several low-level vision tasks: (a) binocular stereo with a slanted plane local model, (b) optical flow with constant flow or affine flow local model and (c) shape from shading with quadratic local model. Reasoning with local models from ``good'' regions (*e.g.* green) will significantly reduce the pixelwise ambiguity in low-level vision, but such models are not universally valid for all regions (*e.g.* red). The key challenge to be addressed in this chapter is how to simultaneously find which regions are valid and estimate the best local model parameters for valid regions.

shapes of the aggregation regions that are right for each part of a scene. Regions that are too small do not sufficiently reduce the underlying ambiguity, while those that are too big or the wrong shape span abrupt scene changes that violate the local model and make estimates unreliable.

We introduce a computational framework to address this challenge. Called the *consensus framework*, we apply it to the binocular stereo problem while also presenting it generally as a way to attack a variety of low-level tasks. The framework explicitly considers a large set of dense, overlapping regions of many sizes that redundantly cover the image plane. It simultaneously determines which regions are inliers to the local model (binary variables) and, for each inlying region, the correct coordinates in the local model space for that region (continuous variables). Estimation is cast as optimizing an objective

that requires each inlying region to be supported by its local image data while also having scene estimates that are consistent with its overlapping neighbor regions. The output of the framework—the inlier statuses of all regions and the local estimates from the inliers—offers a rich, multi-scale representation of the physical scene. This includes spatial grouping information, a global scene map, and a point-wise measure of confidence, all of which are desirable when seeking to combine multiple low-level cues or integrate higher-level processes.

Compared to traditional approaches based on Markov random fields (MRFs), the consensus framework reasons in a much larger variable space, and more critically, with orders of magnitude more links between variables. This is because it enforces simultaneous consistency between the thousands of regions that overlap any single pixel. Despite this complexity, two properties make estimation not only feasible, but efficient. First, since the dense region-set embodies an over-complete scene representation—with many more internal variables than values in the output scene map—good solutions can often be reached by a simple alternating algorithm similar to expectation–maximization (see Section 4.4). Second, when the regions are organized hierarchically by scale, each region only needs to sum information from its parents and children and computation complexity can be significantly reduced (detailed in Section 4.5).

Experiments on the binocular stereo problem show that the consensus framework achieves greater accuracy on the KITTI benchmark [21] than comparable state-of-the-art variational and MRF approaches, as is evaluated in Section 4.7.1. The shape from shading reconstruction algorithm described in the previous chapter (Section 3.5.2) is in fact a preliminary version of the consensus framework proposed in this chapter, and detailed connections and differences will be described in Section 4.7.2.

4.2 Related Work

There are many techniques for low-level vision problems like binocular stereo, optical flow, and shape-from-shading. While they vary greatly in the way they derive information point-wise from image cues,

their mechanisms for spatial aggregation tend to follow one of three different paradigms. The simplest paradigm is purely local—a single support region is explicitly defined around each pixel [37, 70, 93, 81, 90]. These regions are typically determined using intensity and texture information, either independently for each pixel or jointly for all pixels via segmentation, and they succeed when color and texture boundaries are well aligned with boundaries in the latent scene map.

Variational methods form another category. Estimation involves minimizing a per-pixel data cost along with a spatial regularization term that penalizes large derivatives in the scene map [6, 7, 32, 44, 66, 74]. The derivative filters are designed to measure deviations from some implied local model, and the penalty is chosen to promote piecewise adherence while still being convex. Some variational methods employ multi-scale reasoning, through sequential coarse-to-fine optimization [7] or simultaneous penalization of derivatives at multiple scales [2].

The third dominant paradigm are MRF-based methods [47, 77, 84, 87, 88, 89]. These methods explicitly encode piece-wise adherence to the local model (as opposed to the convex penalties in variational methods, which do so implicitly), by making hard decisions about the local model being valid across an edge or clique. Since they often consider continuous label spaces and non-submodular smoothness terms, these methods tend to rely on expensive approximate algorithms for optimization. Computation can be reduced by defining graphs on super-pixels instead of pixels [87, 88, 89], and this does not substantially reduce accuracy as long as the super-pixel boundaries happen to be well aligned with scene boundaries.

The consensus framework is different from traditional, single-scale MRF techniques because it is defined on overlapping regions at multiple scales. It is also different from multi-scale MRF formulations that have been used for segmentation [48], where parent nodes encode semantic context for co-occurring labels of their children. In consensus, all regions at all scales are self-similar. They all make direct predictions about pixel-level scene values, and they all use the same local model.

We use an alternating algorithm to minimize our objective. This is similar to "divide and con-

cur" optimization algorithms like the alternating direction method of multipliers (ADMM) [14] that modify an objective to create multiple copies of a variable—one for each term in the original objective that includes that variable—and then enforce consistency between these copies. Our consensus objective resembles these modified, split objectives. A crucial part of our approach is the hierarchical organization of regions across scales, which makes the aggregation steps in the alternating minimization tractable. It is worth noting the approach of [41] here, which also uses an efficient data-structure for message aggregation during mean-field inference in a densely connected graph.

4.3 Consensus Framework

We begin with a formal description of the three main components of the proposed framework. First, there is the global scene map. This is a function $Z(n) \in \mathbb{R}^d$ on the two-dimensional image plane, with $n = (x, y)$ indexing discrete spatial locations. $Z(n)$ may be scalar-valued ($d = 1$) for properties such as stereo disparity, or vector-valued for properties such as motion and 3D surface orientation. Second, there is a dense set $\mathcal{P}$ of overlapping regions $p \in \mathcal{P}$ within the image plane, each one a collection of locations n. The final component is the local model $Z_p(n; \theta_p)$, where θ_p is the model parameter. The local model is expected to apply piecewise across most of the scene, and it restricts accordingly the allowable choices for scene values within any region p. The proposed framework is very flexible with its mathematical form so as to encompass all sorts of local models proposed in computer vision [4, 86]. A few examples are shown in Figure 4.1, and a concrete family of generalized linear local models will be presented in Section 4.5, which is general enough to include many interesting local models and also facilitates efficient computation at the same time.

With the three components in hand, estimation requires determining: a) which regions $p \in \mathcal{P}$ are inliers with respect to the local model; and b) for all inlying regions, values of the per-region variables θ_p that are supported by the image data and consistent with each other. Inliers are indicated by a binary variable $I_p \in \{0, 1\}$ associated with each patch. Once determined, the values of $\{I_p, \theta_p\}$ together

provide a rich and over-complete representation of the physical scene. At each point n, local grouping information is available through the subset J_n of (potentially thousands of) inlying regions covering that point:

$$J_n = \{p \colon p \ni n, I_p = 1\}. \tag{4.1}$$

An estimate $\bar{Z}$ of the global scene map is induced as the point-wise average, or *consensus*, of the local estimates from inlying regions:

$$\bar{Z}(n) = \frac{1}{|J_n|} \sum_{p \in J_n} Z_p(n; \theta_p) = \frac{1}{\sum_{p \ni n} I_p} \sum_{p \ni n} Z_p(n; \theta_p)\, I_p. \tag{4.2}$$

The count $|J_n|$ represents the *degree of consensus* at each point, and provides a point-wise measure of confidence in the estimate $\bar{Z}$.

Estimation is then cast as a minimization of the following cost over variables $\{I_p, \theta_p\}$:

$$L(\{I_p, \theta_p\}_{p \in \mathcal{P}}) = \sum_{p: I_p = 0} \tau_p + \sum_{p: I_p = 1} D_p(\theta_p) + \lambda \sum_n |J_n| \ \text{Var}\left[\{Z_p(n; \theta_p)\}_{p \in J_n}\right]. \tag{4.3}$$

The first term applies a cost τ_p for declaring region p an outlier, in line with intuition that the local model is often valid. The second term scores local variables θ_p in each inlying region using *data cost* $D_p(\cdot)$, typically measuring the ability of restricted local scene estimates $Z_p(n; \theta_p), \forall n \in p$ to explain the relevant image data. Both τ_p and $D_p(\cdot)$ can optionally be augmented to encode prior information about the scene or context from semantic visual processes. The final λ-weighted term promotes consistency between overlapping regions by penalizing, at every point, the variance of the scene predictions from inlying regions that cover it.

4.4 Alternating Optimization Algorithm

To minimize (4.3), we re-write the consistency term in terms of the global scene map Z, creating a related cost L':

$$L'(\{I_p, \theta_p\}_{p\in\mathcal{P}}, Z) = \sum_{p:I_p=0} \tau_p + \sum_{p:I_p=1} \left(D_p(\theta_p) + \lambda \sum_{n\in p} \|Z_p(n;\theta_p) - Z(n)\|^2 \right), \quad (4.4)$$

where the two costs are equal when Z is set to the consensus as per (4.2), *i.e.* $L'(\{I_p, \theta_p\}, \bar{Z}) = L(\{I_p, \theta_p\})$. We define the *consistency cost* for patch p with respect to a given scene map Z as

$$C_p(\theta_p, Z) = \sum_{n\in p} \|Z_p(n;\theta_p) - Z(n)\|^2. \quad (4.5)$$

Cost L' is minimized iteratively, with each iteration having two steps. The first step is a minimization over region variables $\{I_p, \theta_p\}$ with Z fixed. Conveniently, this can be done independently—and in parallel—for each region since there are no cross-region terms in L' when Z is fixed. These independent minimizations are achieved by setting

$$\theta_p = \arg\min_{\theta} [D_p(\theta) + \lambda C_p(\theta, Z)], \quad (4.6)$$

and then,

$$I_p = \begin{cases} 0, & \text{if } [D_p(\theta_p) + \lambda C_p(\theta_p, Z)] > \tau_p, \\ 1, & \text{otherwise .} \end{cases} \quad (4.7)$$

In other words, the best model-based explanation is found for each region p, and then the region is declared outlier if the error-of-fit exceeds the outlier cost τ_p.

The second step at each iteration is a minimization over Z with region variables fixed at their new values. This is achieved simply by setting $Z = \bar{Z}$ as per (4.2), and it is thus guaranteed that

$L(\{I_p, \theta_p\}) = L'(\{I_p, \theta_p\}, Z)$ at the end of every iteration. Consequently, beginning with any initial estimate of the scene map Z, each iteration decreases the value of L', and therefore of L, which converges to a (local) minimum whenever $\{D_p(\cdot)\}$ have finite lower bounds.

Convergence to a good local minimum is promoted by beginning the iterations with a smaller value for the consistency weight λ, and then increasing it to its final value across the initial iterations. Interestingly, this induces a temporal coarse-to-fine refinement of the scene map during the optimization. Early-on, smaller λ values allow more inlying regions, causing the consensus to be smoothed across larger areas. As λ increases, more regions that span scene discontinuities become outliers, and the consensus exhibits progressively finer detail.

4.5 Hierarchical Computation

The computational cost of the alternating optimization algorithm depends on the complexities of the three parts of every iteration:

1. Computing consistency terms $C_p(\theta_p, Z)$ for every region p as per (4.5).
2. Updating (θ_p, I_p) for every region p as per (4.6) and (4.7).
3. Computing $\bar{Z}$ as per (4.2).

Considering the large number of regions (same order as the number of pixels) and large number of pixels per region (we use regions of size up to 64×64), each of the three steps above can be very expensive. In this section, we show that by organizing patches in a spatial hierarchy and assuming a generalized linear local model, the computation complexity of Step 1 and Step 3 can be significantly reduced. As will be shown in Section 4.6, this hierarchy structure will also benefit Step 2 if the data costs $D_p(\theta_p)$ satisfy certain constraints.

In our hierarchy structure, set of patches $\mathcal{P}$ has regions at K different scales, and symbol $\mathcal{P}_k$ represents the subset of regions at scale $k \in \{1 \dots K\}$. By convention, larger values of k correspond to

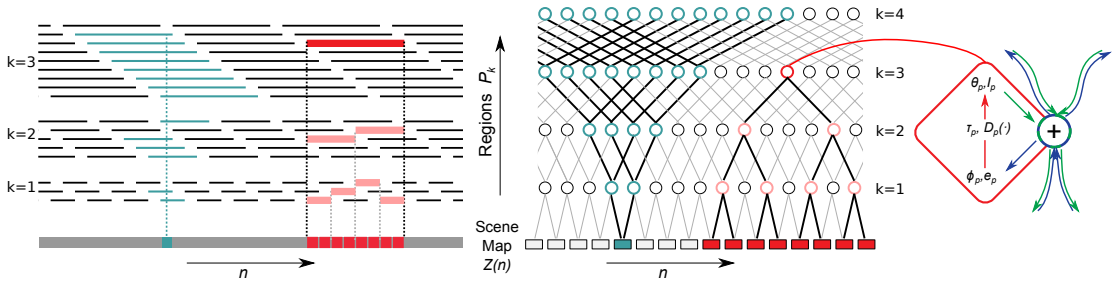


Figure 4.2: Regions are organized hierarchically. Any region $p \in \mathcal{P}_k$ at scale $k > 1$ (say the red one) can be written exactly as the union of a set of non-overlapping child regions from scales smaller than k (the pink ones). This hierarchical structure facilitates efficient computation for consistency term $C_p(\theta_p; Z)$ and scene map $\bar{Z}$.

larger regions. Moreover, the regions can be organized hierarchically: for every region $p \in \mathcal{P}_k$ at scale $k > 1$, it is possible to select a set H_p of non-overlapping "child regions" from scales smaller than k, such that p can be written exactly as their disjoint union. Figure 4.2 shows an example of such a region set for a one-dimensional image plane, where each $\mathcal{P}_k$ is the set of overlapping regions of length 2^k, and each $p \in \mathcal{P}_k, k > 1$ is the union of two children from $\mathcal{P}_{k-1}$.

We restrict the local model to a generalized linear form:

$$Z_p(n; \theta_p) = U(n)\theta_p, \qquad \forall n \in p, \tag{4.8}$$

where $U(n) \in \mathbb{R}^{d \times M}$ is some pre-defined matrix-valued function on the image plane, and $\theta_p \in \mathbb{R}^M$ is a variable associated with region p. Algebraically, this restricts local scene values to an M-dimensional linear subspace, regardless of region size; and as a consequence of using a common $U(n)$, local scene estimates from two overlapping regions p and p' agree whenever $\theta_p = \theta_{p'}$. Here are some

examples of functions $U(n)$ and their corresponding physical interpretations:

$$U(n) = [x \;\; y \;\; 1], \quad d = 1, M = 3, \tag{4.9}$$

(disparity of locally-planar surfaces),

$$U(n) = \begin{bmatrix} \partial/\partial x \\ \partial/\partial y \end{bmatrix} [x^2 \;\; y^2 \;\; xy \;\; x \;\; y], \quad d = 2, M = 5, \tag{4.10}$$

(normals of locally-quadratic surfaces),

$$U(n) = \begin{bmatrix} x \;\; y \;\; 1 & 0 \\ 0 & x \;\; y \;\; 1 \end{bmatrix}, \quad d = 2, M = 6, \tag{4.11}$$

(flow vectors for locally-affine motion).

Under these generalized linear models, the consistency term (4.5) is a quadratic function over model parameters θ_p:

$$C_p(\theta_p, Z) = \sum_{n \in p} \|U(n)\theta_p - Z(n)\|^2 = \theta_p^T Q_p \theta_p - 2\phi_p^T \theta_p + e_p, \tag{4.12}$$

with each $Q_p = \sum_{n \in p} U(n)^T U(n)$ a pre-computed $M \times M$ matrix permanently associated with region p; and each ϕ_p, e_p an M-vector and a scalar, respectively, derived from Z as:

$$\phi_p = \sum_{n \in p} U(n)^T Z(n), \qquad e_p = \sum_{n \in p} \|Z(n)\|^2. \tag{4.13}$$

Using the fact that every region p is partitioned by its child regions $c \in H_p$, we can write

$$\phi_p = \sum_{n \in p} U(n)^T Z(n) = \sum_{c \in H_p} \sum_{n \in c} U(n)^T Z(n) = \sum_{c \in H_p} \phi_c, \tag{4.14}$$

and similarly, $e_p = \sum_{c \in H_p} e_c$. This reduces the number of additions significantly—from the number of pixels in region p to just the number of its children. To ensure that values of $\{\phi_c, e_c\}_{c \in H_p}$ are available, calculations of $\{\phi_p, e_p\}$ are scheduled in an upward sweep through the hierarchy, using explicit summation over pixels for regions at scale $k = 1$, and the cheaper right-most expression of (4.14) for progressively larger scales.

The hierarchical structure can also be leveraged to efficiently compute the consensus $\bar{Z}$ from the current values of the region variables $\{\theta_p, I_p\}$. Note that for every region $p \ni n$ at scale $k > 1$, there is one and only one child region in H_p that also includes n. For the simple case with only two scales ($K = 2$), we see that the summation of local estimates from inlying regions can be simplified to

$$\sum_{\{p \ni n\} \cap \mathcal{P}_1} U(n)\theta_p I_p + \sum_{\{p \ni n\} \cap \mathcal{P}_2} U(n)\theta_p I_p = \sum_{\{p \ni n\} \cap \mathcal{P}_1} U(n) \left(\theta_p I_p + \sum_{r \in H_p^{-1}} \theta_r I_r \right), \quad (4.15)$$

where $H_p^{-1} = \{r : H_r \ni p\}$ denotes the set of *parents* for any region p.[1] In the more general case with K scales, we recursively define augmented variables $\{\theta_p^+, I_p^+\}$ for every region p as

$$\theta_p^+ = \theta_p \, I_p + \sum_{r \in H_p^{-1}} \theta_r^+, \quad I_p^+ = I_p + \sum_{r \in H_p^{-1}} I_r^+, \quad (4.16)$$

which can be computed by a *downward* sweep through the pyramid. Then, it is easy to see that the numerator and denominator of the expression for $\bar{Z}(n)$ in (4.2) are given by

$$\left(\sum_{p \ni n} U(n)\theta_p \, I_p \right) = U(n) \sum_{\{p \ni n\} \cap \mathcal{P}_1} \theta_p^+, \quad (4.17)$$

$$\left(\sum_{p \ni n} I_p \right) = \sum_{\{p \ni n\} \cap \mathcal{P}_1} I_p^+. \quad (4.18)$$

[1] Note that for a region $p \in \mathcal{P}_K$ at the largest scale, $H_p^{-1} = \emptyset$.

Thus, instead of computing summations over all overlapping regions at all scales for each location n, the consensus can be computed using summations over the augmented variables $\{\theta_p^+, I_p^+\}$ of regions at just the smallest scale.

The gains from using these recursive computations is substantial, and can be interpreted as reducing the *effective* connectivity of the framework to just the sparse set of hierarchical links. For the network in Figure 4.2, it represents a reduction, in the number of required summations for (4.2) and (4.5), from $\mathcal{O}(2^K N)$ to $\mathcal{O}(KN)$. Moreover, while the recursion requires different scales to be processed sequentially, note that the computations in (4.14) and (4.16) can still be carried out for all regions $p \in \mathcal{P}_k$ at each scale k in parallel. Therefore, as visualized in right most plot of Figure 4.2, computation happens in a distributed architecture, requires the identical operations of (4.6), (4.7), (4.14), and (4.16) at each region, with operations at each scale happening in parallel and information being passed through hierarchical links between scales—all of which arises naturally as an efficient way to optimize a well-defined mathematical objective.

4.6 Data Costs in Binocular Stereo

As in many low-level vision frameworks, one of the key ingredients is a proper data cost function. In order to apply the proposed consensus framework to a specific vision task, one needs to provide a data cost function $D_p(\theta)$ for each local region p, and decide an appropriate optimization method to minimize the cost function $D_p(\theta) + \lambda C_p(\theta, Z)$ in (4.6) for the alternating optimization algorithm, where $C_p(\theta, Z)$ is the consistency term defined in (4.5), and often quadratic of θ as per (4.12) if assuming a generalized linear local model. This minimization needs to be carried out on every region—independently and therefore can be done in parallel—and in every iteration, which accounts for most of computation in the optimization algorithm. In this section, we describe the representation and minimization for data cost in the context of a binocular stereo matching application. Many of the techniques are relevant to other low-level applications such as shape from shading and optical flow.

We first define a pixel-based data cost volume $V(n, z)$, where $n = (x, y) \in \mathbb{Z}^2$ denotes a pixel location in the visual field (which we assume is equal to the left image, as is common in binocular stereo) and $z \in \mathbb{Z}$ denotes the discrete disparity. This cost volume is also known as disparity space image (DSI), *e.g.* in [71], whose value $V(n, z)$ describes how well the pixel $n = (x, y)$ in the left view image matches the corresponding pixel $n' = (x + z, y)$ in the right view.[2] There are many methods for robustly computing this cost volume, including subpixel-sampled absolute difference [3], Hamming distance of census transform [91], mutual information based cost followed by semi-global matching [30] and learned cost functions using a convolutional neural network [92]. Unless explicitly stated, the methods described in this section assume a cost volume has been pre-computed and supplied as input, but do not care which specific algorithm was used to compute it.

Given a pre-computed pixel-based cost volume $V(n, z)$, the data cost $D_p(\theta)$ for each patch p can be simply calculated as the accumulation of costs for all pixels in the patch with corresponding disparity predicted by the local model $Z_p(n; \theta) = U(n)\theta$. More specifically, the data cost is defined as

$$D_p(\theta) = \sum_{n \in p} V(n, U(n)\theta). \tag{4.19}$$

where $U(n)\theta$ gives the disparity of pixel n predicted by the model θ. We assume $U(n)$ to be the slanted plane model defined in (4.9) for the rest of this chapter, but one can use other choices (*e.g.* frontal parallel or locally quadratic models) as well.

Defining the region data cost as the accumulation of per-pixel costs as per (4.19) facilitates efficient computation when regions are organized hierarchically. This is because data costs of the form (4.19) naturally imply $D_p(\theta) = \sum_i D_{c_i}(\theta)$, where region p is the union of non-overlapping child regions,

[2] Note that some methods compute this cost volume by comparing matches not for a single pixel, but a small patch centered at that pixel instead. This is equivalently to making a frontal parallel assumption for a small enough neighborhood around the pixel, and use the matching quality of the neighborhood as a proxy for that of the centering pixel.

i.e. $p = \cup_i c_i$ and $c_i \cap c_j = \emptyset,\ \forall i \neq j$.

The challenge of working with data costs in stereo (and other low-level vision problems) is that the data cost functions are usually complicated, *e.g.* non-convex, flat in a large region , contains many local minima, *etc.* These complication makes the local minimization (4.6) inefficient to perform, and prone to getting trapped in poor local minima. In the rest of this section, we describe two approaches to address this difficulty.

4.6.1 Tabulated Cost Function

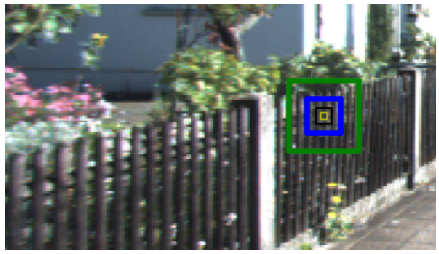
One natural way of representing and minimizing a complicated cost function $D_p(\theta)$ is to tabulate it, or in other words, densely sample it on a regular grid $\Theta = \{\theta^{(j)}\}_j$. [3] Denote the set of valid slanted planes—the space of valid θ—as $\Xi \subset \mathbb{R}^3$. In this section, we first derive a finite boundary for Ξ, and then find a grid step size such that any valid $\theta \in \Xi$ is sufficiently close to a sample on the grid $\theta^{(j)} \in \Theta$. Finally, we show that the tabulated grid for a region $p \in \mathcal{P}_k$ can be directly computed as a transformed-sum from grids of its non-overlapping children $H_p \subset \mathcal{P}_{k-1}$ (see Figure 4.2).

We introduce local coordinates $n_p = (x_p, y_p)$, because, as will be shown shortly, this allows uniform quantization of valid θ space Ξ for all regions, regardless of their global locations in the image. A region of size $A \times A$ has its local origin $(0, 0)$ at top-left corner, and its local coordinates $(x_p, y_p) \in \Omega_p = [0, A] \times [0, A]$. The slanted plane local model is given as

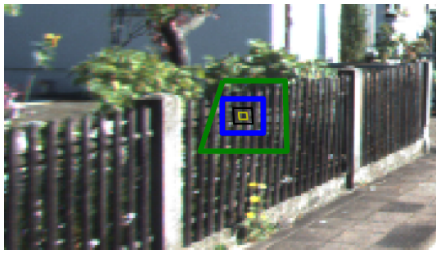
$$Z(n_p; \theta) = U(n_p)\theta = \theta_x x_p + \theta_y y_p + \theta_0. \tag{4.20}$$

Before diving into the detailed analysis, we first show some example tabulated cost volumes in Figure 4.3. The selected regions are centered at the same pixel but of various different scales. They contain

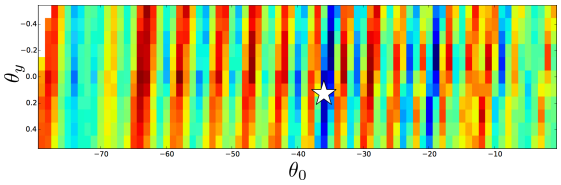
[3] When describing different local model parameters θ, we use subscript such as θ_p to indicate the model parameter designated to a region p, and superscript such as $\theta^{(j)}$ to indicate a sample in the parameter space. We also use subscript to denote individual element of the $\theta \in \mathbb{R}^3$ vector, written as θ_x, θ_y and θ_0, and when appearing together with region subscript p, we put a colon in the middle and write $\theta_{p:x}$, $\theta_{p:y}$ and $\theta_{p:0}$.



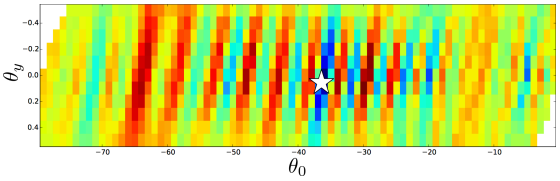
(a) Left View.



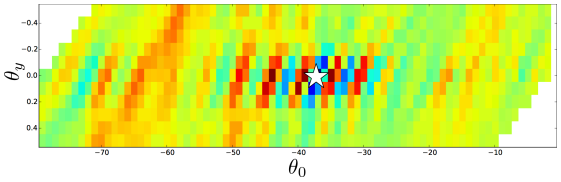
(b) Right View.



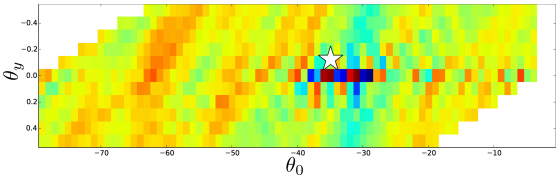
(c) Tabulated cost for a 4×4 region.



(d) Tabulated cost for a 8×8 region.



(e) Tabulated cost for a 16×16 region.



(f) Tabulated cost for a 32×32 region.

Figure 4.3: Tabulated cost function for regions centered at the same pixel but of different scales. We show one slice of the 3D cost volume $D(\theta_x, \theta_y, \theta_0)$, minimizing the θ_x dimension out. In other words, the plots (c)-(f) visualize the value of function $f(\theta_y, \theta_0) = \min_{\theta_x} D(\theta_x, \theta_y, \theta_0)$. The white star in each plot indicates the ground truth parameter.

repetitive textures, and therefore the cost volumes for small regions (4×4 and 8×8) show periodic patterns, and based on them we are not able to resolve which region on the other view best matches the current one. When the region gets bigger (16×16), enough contextual information is gather and a distinct minima emerges. As the region size keeps growing (32×32), it will begin to cross disparity discontinuity, and the cost volume provides erroneous information (even though a prominent minima still exists, it is not consistent to the region's neighbors and therefore the consensus framework will likely to declare the region an outlier).

Bounds for valid model parameters

Let $Z_{\min}, Z_{\max}$ be the minimum and maximum disparity of the scene, then for any valid slanted plane $\theta \in \Xi$, we have

$$Z_{\min} \leq Z(n_p; \theta) \leq Z_{\max}, \quad \forall n_p \in \Omega = [0, A] \times [0, A]. \tag{4.21}$$

The minimum and maximum disparity of a planar region happens at the extreme points, and therefore for a plane θ to be valid, we simply need to check

$$Z_{\min} \leq Z(n_p; \theta) \leq Z_{\max}, \quad \text{for } n_p \in \{(0,0), (0,A), (A,0), (A,A)\}. \tag{4.22}$$

Note that this is a polyhedron with eight faces in $\mathbb{R}^3$. Next we find a cube

$$\Xi^* = \left[\theta_x^{\min}, \theta_x^{\max}\right] \times \left[\theta_y^{\min}, \theta_y^{\max}\right] \times \left[\theta_0^{\min}, \theta_0^{\max}\right] \subset \mathbb{R}^3 \tag{4.23}$$

that tightly bounds this polyhedron. It is easy to see that we can set $\theta_0^{\min} = Z_{\min}$ and $\theta_0^{\max} = Z_{\max}$. For $\theta_x^{\min}$ and $\theta_x^{\max}$, the tightest bound can be found by the fact that

$$|\theta_x A| = |Z(A, y) - Z(0, y)| \leq Z_{\max} - Z_{\min}, \quad \Longrightarrow \quad |\theta_x| \leq \frac{Z_{\max} - Z_{\min}}{A}. \tag{4.24}$$

The same bounds apply for θ_y.

Therefore we found $\Xi^* = \{\theta \in \mathbb{R}^3 : -\frac{Z_{\max}-Z_{\min}}{A} \leq |\theta_x|, |\theta_y| \leq \frac{Z_{\max}-Z_{\min}}{A}, Z_{\min} \leq \theta_0 \leq Z_{\max}\}$ as the cube that most tightly bounds the valid θ space Ξ.

Grid step size

Next we need to determine a grid step size, such that any valid $\theta \in \Xi$ is sufficiently close to a sample on the grid. For two slanted planes θ and θ', we define their ℓ_∞ distance as the maximum disparity difference between them

$$\ell_\infty(\theta, \theta') = \max_{n_p \in \Omega_p} \left| Z(n_p; \theta) - Z(n_p; \theta') \right|. \tag{4.25}$$

Following [40], we define an *ϵ-net* as a set $\Theta \subset \mathbb{R}^3$ such that for any valid $\theta \in \Xi$, there exists a $\theta^{(j)} \in \Theta$ whose distance to θ is not larger than ϵ, *i.e.*

$$\forall \theta \in \Xi,\ \exists \theta^{(j)} \in \Theta,\ \text{ such that } \ell_\infty(\theta, \theta^{(j)}) \leq \epsilon. \tag{4.26}$$

For two slanted planes θ and θ', as long as the differences of their θ_x and θ_y components are less than $2\delta/A$ and the difference of their θ_0 components is less than δ, we can guarantee that their ℓ_∞ distance is less than 3δ, because

$$\begin{aligned}
&\left|(\theta_x x_p + \theta_y y_p + \theta_0) - \left(\theta'_x x_p + \theta'_y y_p + \theta'_0\right)\right| \\
\leq &|x_p| \cdot \left|\theta_x - \theta'_x\right| + |y_p| \cdot \left|\theta_y - \theta'_y\right| + \left|\theta_0 - \theta'_0\right| \\
\leq &\frac{A}{2} \cdot \frac{2\delta}{A} + \frac{A}{2} \cdot \frac{2\delta}{A} + \delta \\
= &3\delta, \qquad \forall (x_p, y_p) \in \Omega_p.
\end{aligned}$$

This imples that in order to achieve a $\frac{3\delta}{2}$-net, we need to sample at step size

$$\Delta\theta_x = \Delta\theta_y = \frac{2\delta}{A}, \quad \Delta\theta_0 = \delta. \tag{4.27}$$

Assume we take $B+1$ samples in the θ_0 dimesion, and suppose we index each sample with a tuple of three integers $j = (j_x, j_y, j_0) \in \mathbb{Z}^3$. Given an index j, the coordinates of the plane $\theta^{(j)}$ can be calculated as

$$\theta^{(j)} = \left(j_x \frac{2\delta}{A},\ j_y \frac{2\delta}{A},\ \frac{Z_{\min} + Z_{\max}}{2} + j_0 \delta \right), \quad \text{with } \delta = \frac{Z_{\max} - Z_{\min}}{B+1}. \tag{4.28}$$

The sampling index space is

$$\begin{aligned} j_x &\in \left[-\frac{\theta_x^{\max} A}{2\delta} - \frac{1}{2},\ \frac{\theta_x^{\max} A}{2\delta} + \frac{1}{2} \right] \cap \mathbb{Z}, \\ j_y &\in \left[-\frac{\theta_y^{\max} A}{2\delta} - \frac{1}{2},\ \frac{\theta_y^{\max} A}{2\delta} + \frac{1}{2} \right] \cap \mathbb{Z}, \\ j_0 &\in \left[-\frac{B}{2}, \frac{B}{2} \right] \cap \mathbb{Z}. \end{aligned} \tag{4.29}$$

This sampling scheme creates a $\frac{3\delta}{2}$-net in Ξ^*, which is a tight superset of valid slanted planes Ξ.

Hierarchical computation of the data costs

Evaluating $D_p(\theta^{(j)})$ cost function at large number of samples $\theta^{(j)}$ for large regions can be very expensive. Fortunately, the spatial hierarchy structure of the framework implies that one does not need to aggregate the data cost of a region from all the pixels $n \in p$ directly; instead, it can be calculated as the sum of its fixed number of child regions $c \in H_p$, reducing the computation complexity from exponential (with respect to region level k) to constant (the size of set H_p, which is 4 for two-dimensional regions). In our tabulation scheme, we need to make sure that the samples needed from region $p \in \mathcal{P}_k$ have already been calculated in the regions of $H_p \subset \mathcal{P}_{k-1}$.

Assume we have a region $p \in \mathcal{P}_k$ and one of its child regions $c \in H_p$, and the size of child region is half the size of its parent, *i.e.* $A_c = \frac{1}{2} A_p$. [4] The region centers are offset by $(\Delta x, \Delta y) \in$

[4] We assume all regions are square in this chapter, but the analysis can also be extended to non-square rect-

$\{(0,0),(0,A_c),(A_c,0),(A_c,A_c)\}$, and we consider the bottom right child as an example (the calculation directly applies to the rest)

$$x_p = x_c + A_c, \quad y_p = y_c + A_c. \tag{4.30}$$

The disparity region model of the parent region is

$$Z(x_p, y_p; \theta_p) = \theta_{p:x} x_p + \theta_{p:y} y_p + \theta_{p:0}, \tag{4.31}$$

which can be written in child region's local coordinates as

$$\begin{aligned} Z(x_c, y_c; \theta^c) &= \theta_{p:x}(x_c + A_c) + \theta_{p:y}(y_c + A_c) + \theta_{p:0} \\ &= \theta_{p:x} x_c + \theta_{p:y} y_c + A_c \left(\theta_{p:x} + \theta_{p:y}\right) + \theta_{p:0}. \end{aligned} \tag{4.32}$$

This means we can transform the model parameter as

$$(\theta_{c:x}, \theta_{c:y}, \theta_{c:0}) = \left(\theta_{p:x}, \theta_{p:y}, A_c \left(\theta_{p:x} + \theta_{p:y}\right) + \theta_{p:0}\right). \tag{4.33}$$

In the index space, assume we have resolution δ in parent level, which gives

$$\theta_{p:x} = j_{p:x} \frac{2\delta}{A^p} = j_{p:x} \frac{\delta}{A_c}, \quad \theta_{p:y} = j_{p:y} \frac{2\delta}{A^p} = j_{p:y} \frac{\delta}{A_c}, \quad \theta_{p:0} = \frac{Z_{\min} + Z_{\max}}{2} + j_{p:0}\delta. \tag{4.34}$$

This means if we sample the child region at resolution $\frac{\delta}{2}$, we will get all the samples needed from θ_c to

anglar regions, as long as the length of each dimension of the child region is either a half or the same as the length of its parent region in the same dimension.

compute θ_p, and the index of the parent region can directly transfer to that of its child as

$$j_{p:x} \to j_{c:x}, \quad j_{p:y} \to j_{c:y}, \quad j_{p:0} \to 2j_{c:0} + (j_{c:x} + j_{c:y}). \tag{4.35}$$

4.6.2 Quadratic Data Cost Based on Semi-Global Matching

The tabulation method in the previous section describes a general approach for addressing the complication of the data cost $D(\theta)$. However, even with the efficient hierarchical computation, the method is still very expensive, because the Θ grid usually contains thousands of samples in order to cover the full parameter space. In this section, we proposed a more stereo-specific cost function based on semi-global matching (SGM) [30] and applied a quadratic approximation to reduce computation complexity. This data cost is widely adopted in modern MRF-based stereo methods such as [87, 88, 89], *etc*.

Following [87], we first combine the gradients-based data cost (absolute difference of gradient input image) and the Hamming distance of Census transform [91] as the raw cost volume, and then apply semi-global block matching algorithm proposed in [30], during which lef-right consistency is checked by computing both left and right disparity maps. This will give us an initial set of approximate disparity estimates $Z_{\text{SGM}}(n)$ at a semi-dense set of locations $n \in \Omega_{\text{SGM}}$. The data costs for every region p are then defined as:

$$D_p(\theta) = \sum_{n \in p} w_{\text{SGM}}(n) \Big(U(n)\theta - Z_{\text{SGM}}(n) \Big)^2, \tag{4.36}$$

where $w_{\text{SGM}}(n) = 0$ if $n \notin \Omega_{\text{SGM}}$, $1/4$ if there is a discontinuity in Z_{SGM} around n, and 1 otherwise.

Note that the data cost defined in (4.36) is a quadratic function of θ. Since the consistency cost $C_p(\theta, Z)$ is also quadratic as per (4.12), their linear combination $D_p(\theta) + \lambda C_p(\theta)$ is quadratic over θ as well, and its minimization can be carried out with a simple 3×3 Cholesky decomposition, which is independent to patch size. Furthermore, the aggregation of children data costs to get that of parent

can also be efficiently done by adding the quadratic coefficients together, which again is independent to patch size. Therefore, by making a quadratic approximation on the local data model in (4.36) will significantly reduce computation complexity of the alternating algorithm, and as will be shown in the experiment section, this approximation does not lose much accuracy because the powerful semi-global matching step has already robustly cleaned up the cost volume.

4.7 Experiments and Evaluation

4.7.1 Binocular Stereo

We first evaluate the proposed framework using SGM-based quadratic data costs described in Section 4.6.2 on the KITTI benchmark [21]. The KITTI dataset contains a total of 389 grayscale image pairs of rural road scenes, captured using an autonomous driving platform equipped with a pair of high-resolution cameras. A Velodyne laser scanner provides ground truth at a subset of pixels in each scene. This ground truth is made available for a subset of 194 image pairs—the *training set*—and withheld for the remaining image pairs that form the *testing set*. A website associated with the database tracks the performance of stereo algorithms on the testing set. Note that while the benchmark also contains temporally-adjacent stereo frames that allow simultaneous reasoning about optical flow and stereo, we ignore those extra frames and consider the pure stereo problem here.

Figure 4.4 visualizes various aspects of the internal representation of our framework on convergence, for three scenes in the KITTI training set. The top row shows the consensus global disparity map, and Rows 2–6 visualizes a regularly-spaced subset of in the inlier statuses I_p. Row 7 provides another view of variables I_p, by explicitly showing some of the "support regions" formed as the union of all patches in J_n, for various pixels n. These regions by-and-large group together points whose disparity values would be well-explained by a slanted plane model. As expected, there is significant variation in the size and shapes of the support regions across each scene, matching the scale of the underlying

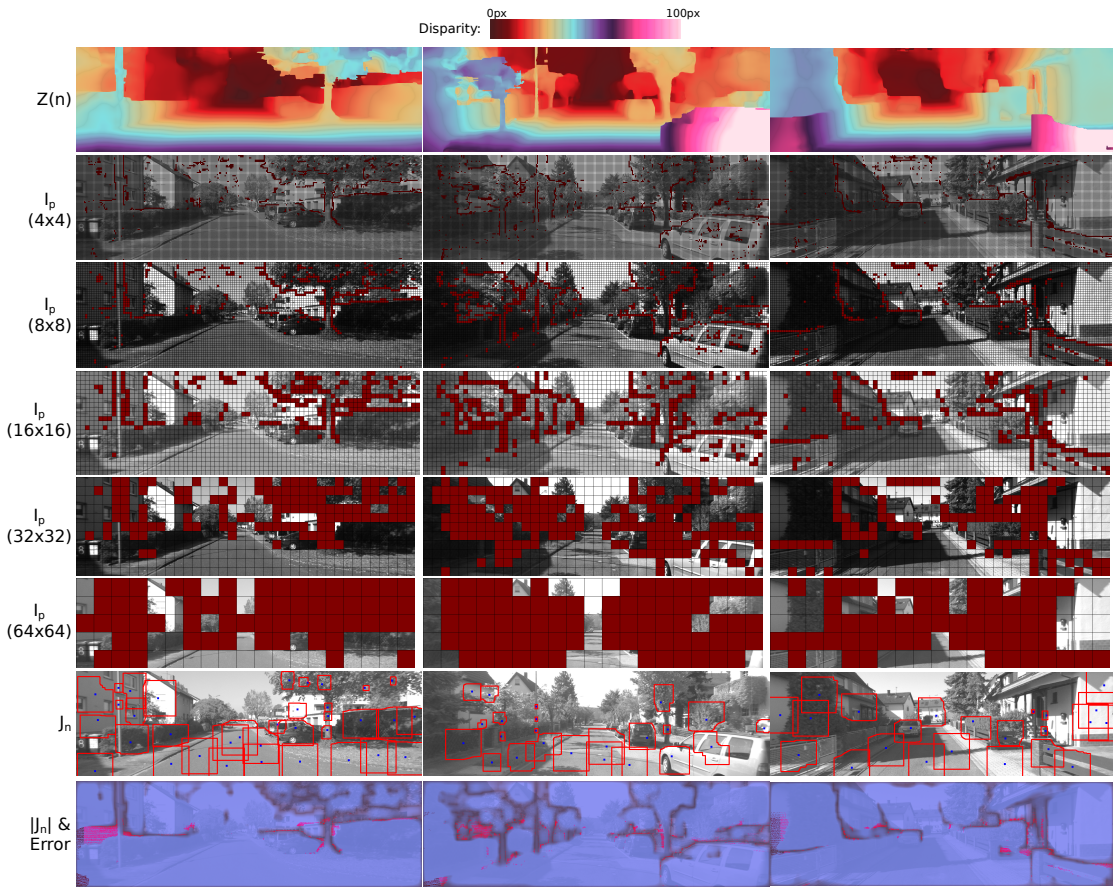


Figure 4.4: Framework output for three image pairs from the KITTI training set. (Row 1) Scene map formed through consensus of predictions from all inlying regions. (Row 2-6) Inlier statuses of regions at different scales, superimposed on the left images of each stereo pair. For clarity, we only show the statuses of a non-overlapping subset of regions at each scale. (Row 7) Boundaries (in red) of the support region for various points n (in blue), formed as the union of their inlying consensus set J_n. (Row 8) Degree of consensus $|J_n|$ (blue saturation) and sites of erroneous estimates (red), defined as estimates with error greater than 3 pixels. (Erroneous estimates whose ground truth disparities place them outside the field of view of the right camera are shown as dark red.)

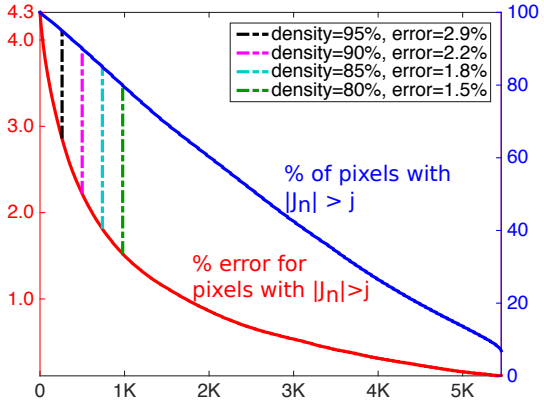


Figure 4.5: Error vs. degree of consensus $|J_n|$. Blue curve shows percentage of points with $|J_n|$ above different thresholds, and red curve their corresponding error rate, in terms of percentage with error > 3 px. These are computed over all pixels with ground truth data available, across all images in the KITTI training set.

scene structures. This highlights the distinction from superpixel-based MRF approaches [87, 88, 89], which require choosing a single scale for the entire scene. Also note that for many pairs of points that do not directly lie in each others' support regions, the regions themselves have significant overlap. Through such overlap, the consensus estimate at a point has benefited from aggregation across regions that are larger than the union of the set of patches that include it.

The final row in Figure 4.4 visualizes the degree of consensus $|J_n|$ at all points (blue saturation), simultaneously with locations of erroneous estimates (red). We see that many of the errors occur around object boundaries and near small scene structures, which are also points where $|J_n|$ is low. We quantify this observation in Figure 4.5, and find that average estimation error drops rapidly as we discard points with the lowest values of $|J_n|$. This another benefit of the rich internal representation: in addition to providing a global scene estimate, it also provides a natural measure of point-wise confidence in this estimate.

Table 4.1 compares the consensus framework with other state-of-the-art stereo algorithms[5] in terms

[5] This table was extracted from the official website on early 2015 and only includes methods that use just one stereo pair as input.

Method	Avg. Error		> 2px		> 3px		> 4px		> 5px		Exec. Time
	All	NOC	All	NOC	All	NOC	All	NOC	All	NOC	
ATGV [66]	1.6px	1.0px	9.05%	7.08%	6.88%	5.02%	5.76%	3.99%	5.01%	3.33%	6min: 8 cores
wSGM [75]	1.6px	1.3px	8.72%	7.27%	6.18%	4.97%	4.89%	3.88%	4.11%	3.25%	6s: 1 core
AARBM [17]	1.2px	1.0px	8.70%	7.36%	5.94%	4.86%	4.56%	3.67%	3.69%	2.96%	0.25s: 1 core
*PCBP [87]	1.1px	0.9px	7.62%	5.08%	5.37%	4.04%	4.29%	3.14%	3.64%	2.64%	5min: 4 cores
*StereoSLIC [88]	1.0px	0.9px	7.20%	5.76%	5.11%	3.92%	4.04%	3.04%	3.33%	2.49%	2.3s: 1 core
*DDS-SS [83]	1.0px	0.9px	6.96%	5.91%	4.59%	3.83%	3.49%	2.90%	2.83%	2.36%	1min: 1 core
*PCBP-SS [88]	1.0px	0.8px	6.75%	5.19%	4.72%	3.40%	3.75%	2.62%	3.15%	2.18%	5min: 1 core
*SPS-St [89]	1.0px	0.9px	6.28%	4.98%	4.41%	3.39%	3.52%	2.72%	3.00%	2.33%	2s: 1 core
MC-CNN [92]	1.0px	0.8px	5.39%	4.30%	3.84%	2.61%	3.01%	2.04%	2.52%	1.75%	100s: GPU
*Proposed: All n	0.9px	0.8px	5.88%	4.85%	4.10%	3.30%	3.26%	2.59%	2.74%	2.16%	6s: 6 cores
Only $\|J_n\| \geq 200$ (96.4% density)	0.8px	0.6px	4.59%	3.50%	2.98%	2.14%	2.26%	1.56%	1.85%	1.24%	

Lowest Second Lowest Third Lowest *: Same Matching Cost

Table 4.1: Comparison with the state-of-the-art on the KITTI testing set. Performance is measured in terms of average error, as well as percentage of estimates with error greater than different thresholds. For each metric, the ``All'' column reports values computed over all ground truth pixels, and ``NOC'' over only those those that are within the field-of-view of the right camera. The last row reports the accuracy of our method's estimates that have confidence measure $|J_n|$ above a threshold, and correspond to errors values computed over 96.4% of the points with ground truth available.

of various error quantiles on the KITTI testing set. The most direct comparisons of our results are with those of [83, 87, 88, 89], since these methods all use the same approach to derive their data costs (census transform and gradient-based matching with SGM). These only differ—from us, and from each other—in their approach to spatial aggregation. The consensus framework outperforms all of these methods on all error metrics, while also having a low execution time.

Table 4.1 also reports the performance of the MC-CNN [92] algorithm, which computes point-wise matching costs using a multi-layer convolutional neural network. This produces lower error values than all other methods, including ours, in exchange for greater computation (the method takes 100 seconds on a GPU with 2880 CUDA cores). This is encouraging, because improved pixel-wise data costs like this one can be directly substituted into the consensus framework to enhance accuracy.

We demonstrate the benefit of the pixel-wise confidence measure in our framework by reporting a second set of results in Table 4.1. This is simply produced by discarding a small number of pixels with

Method	Grid step size	$>$ 3px NOC	$>$ 4px NOC	Exec. Time
Quadratic approximation of SGM costs	N/A	4.16%	3.28%	6s: 6cores
Quadratic approximation of census transform costs	N/A	20.77%	17.31 %	6s: 6cores
Tabulated SGM costs	1	4.11%	3.30 %	41s: 8cores
Tabulated SGM costs	2	4.23%	3.47 %	13s: 8cores
Tabulated SGM costs	4	4.71%	3.93 %	3s: 8cores
Tabulated census transform costs	1	8.25%	7.27 %	41s: 8cores
Tabulated census transform costs	2	8.97%	8.13 %	13s: 8cores
Tabulated census transform costs	4	10.01%	8.68 %	3s: 8cores

Table 4.2: Comparison of tabulated cost functions with the quadratic-approximated cost functions, using SGM-processed data costs or raw census transform data costs. The evaluation is done on a subset of 20 image pairs in the training set.

the lowest degree of consensus $|J_n|$ (*i.e.* those with degree less than 200 out of the maximum possible value of ~ 5500). This second set of error quantiles—computed now on the high-confidence set with only $\sim 3.6\%$ fewer pixels—are the smallest of all methods. In this mode, the proposed algorithm efficiently produces very reliable disparity estimates, at all but a small fraction of locations. This also suggests a strategy for leveraging sophisticated matching strategies such as [92] when execution time is a bottleneck (such as for automated driving applications)—one where the more expensive matching costs are computed only for the small number of low-confidence pixels.

Finally, we test the tabulated data costs method described in Section 4.6.1, and also try the raw data costs without SGM processing. The evaluation results are shown in Table 4.2. We observe that when using SGM-processed data costs, the tabulation approach achieves similar accuracy as quadratic approximation approach if the grid step size is small enough, and that the performance degenerates gracefully as the step size increases, with execution time decreasing quadratically. We also found that algorithms using raw census transform data costs produce results inferior to those using SGM-processed data costs. This is because without the weak spatial filtering and left-right consistency check provided by SGM, the raw data costs are very noisy and erroneous, leading to reduced accuracy in the final estimates. This is particularly evident from the large errors of the quadratic approximation of the raw costs (second row of Table 4.2), because the noisy raw costs are poorly approximated by the smooth quadratic functions. The results suggest that in addition to the choice of inference framework, the

quality of data costs also plays a crucial role in the final reconstruction. For stereo application, the well-established SGM algorithm is especially beneficial for correcting erroneous matches in the raw data cost volume and thus providing a much more robust starting point for spatial reasoning with the inference framework.

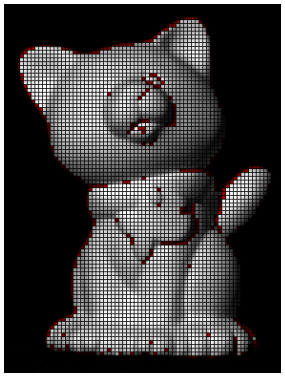
4.7.2 Shape from Shading

The shape from shading algorithm in the previous chapter (Section 3.5.2, detailed in [86]) is in fact a preliminary version of the proposed consensus framework. The algorithm assumes a quadratic local model and does inference on a set of overlapping regions at different scales. For each region, the algorithm simultaneously determines whether it is an inlier or outlier, and for inliers, which shape proposal to take. The algorithm also enforces the estimates of inlying regions to be consistent and requires the result normal vector field able to form a continuous surface through an integrability constraint. Finally, a cooling schedule is present in the algorithm which induces a temporal coarse-to-fine refinement and helps the alternating optimization to avoid being stuck at poor local minima.

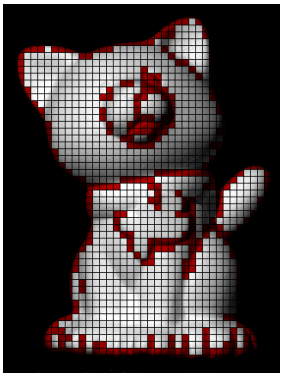
There are also several differences. In the shape from shading reconstruction, we first compute a set of possible shape proposals from a given shading region, and then fix that set in the following iterations. This is equivalent to say the possible θ space Ξ is restricted to a finite set during alternate optimization. The reason we did this is because by the time we developed the shape from shading algorithm, the computationally efficient hierarchical structure was not discovered and performing continuous optimization to find best θ in $\mathbb{R}^5$ for every region in every iteration was formidably expensive. Another difference is that when aggregating from local models to global normal field and depth map, we do not simply take the average, but use a normal-to-depth integration algorithm [19]. The reason for this step is that for this particular application, we have additional knowledge about the physical scene property (normal vector field needs to be integrable).

Figure 4.6 visualizes the inlier maps I_p output by the algorithm at different scales. We observe that

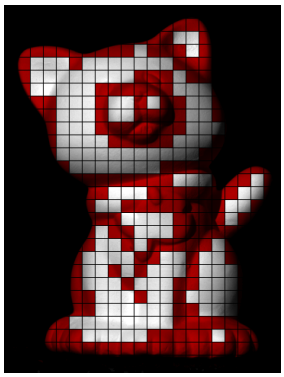
(a) Input image



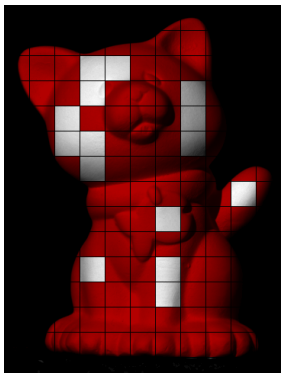
(b) Inlier map I_p at 5×5 scale



(c) Inlier map I_p at 9×9 scale



(d) Inlier map I_p at 17×17 scale



(e) Inlier map I_p at 33×33 scale

(f) Confidence map $|J_n|$

Figure 4.6: Similar to Figure 4.4, we show the inlier and confidence maps output by our framework for the shape from shading application. Red squares in subplots (b)-(e) indicate those regions are inferred as outliers, which are usually caused by discontinuities in normal vector field.

the detected outlier regions usually stride the discontinuity boundaries of the normal vector field as expected. Figure 4.4 (f) shows the pixelwise confidence map $|J_n|$, which has high values at smooth regions because there are more inliers around the center pixel that provide reliable local estimates.

4.8 Conclusion

In this chapter, we introduced a framework for low-level vision with local scene models that reasons with a large overlapping, multi-scale set of regions, to determine which of them are outliers, and which of them can generate model-based scene value estimates while being consistent with each other. Despite the larger variable space, and the greater complexity of the consensus objective, we showed that optimization can be carried out efficiently by recognizing that the regions can be organized hierarchically. We presented an example application of the framework to the stereo matching problem, and showed two different approaches on modeling and minimizing the data cost. An evaluation on stereo estimation found that the framework outperforms existing approaches to spatial reasoning. We also discussed the connection and differences of the proposed framework to the shape from shading reconstruction algorithm presented in the previous chapter.

5

Discussion and Future Work

This dissertation presents theory and algorithms for physics-based visual inference, which aims to recover scene properties from images. Visual inference is the inverse of the image formation process, and is almost always ill-posed because information is inevitably lost when images are formed. This dissertation addresses the information loss in both steps during image formation: the physical process where light interacts with objects in the scene according to their geometry and material properties, and the measurement process where consumer digital cameras typically distort the radiometric signal in order to produce visually pleasing images. We explicitly account for the inherent uncertainties during inference, by reducing them with local models and reporting them to downstream applications.

In Chapter 2, we advocate the use of probabilistic approaches over deterministic ones for color de-rendering, showing that probabilistic de-rendering embraces the multivalued nature of the rendered-

color-to-scene-color map and therefore does not require discarding any image data using ad-hoc thresholds, which is usually necessary in deterministic de-rendering. There are still several unexplored directions that can further improve our ability to infer linear scene colors from distorted sRGB images:

- Currently we use non-parametric models such as Gaussian process or support vector regression, which can accurately adapt to the statistical nature of the data, but on the other hand is inefficient in terms of usage and storage. Because of the large number of implicit parameters (*i.e*. the underlying training data), the models are usually slow to apply on test data and take lots of space in memory or on disk. This inefficiency gets even worse as the amount of training data grows. It will be useful to develop compact representations for probabilistic de-rendering.

- There are several effects in the color processing pipeline that have not been fully examined yet. One example is the (auto-)white balance module that is widely available in today's digital cameras. Although a linear process, it can significantly affect the gain factors of one or more color channels and hence changes the overall forward mapping. It is worth digging in to see whether reliable white balance parameters are available from the output metadata (*e.g*. EXIF tags), or dedicating an additional step to estimate them from the compact images using natural image statistics.

- Another shortcoming of the proposed de-rendering approach is the requirement of an expensive offline calibration process, which needs collection of RAW/JPEG pairs on a variety of different scene colors. In practice, this is possible only by artificially creating extreme illuminations (*e.g*. using gel filters). The process also needs to be repeated to every different imaging mode of every camera. Such burdens could possibly be reduced by studying and understanding the characteristics of forward and backward mappings, which might enable calibration by extrapolation with fewer data points, and transferring the calibrated model from one imaging mode to another imaging mode of the same camera, or even from one camera to another camera.

In Chapter 3, we present mathematical analysis showing that by assuming a quadratic local model, the ambiguities of shape from shading can be dramatically reduced from a continuous manifold to a small discrete set. There are a few theoretical questions remains to be answered:

- The analysis is performed in two essentially different setups—extrinsic and intrinsic—and the derived uniqueness properties are qualitatively different. Two things change between these setups: the geometric families of local patches are different because of the change of coordinate systems, and the imaging models are different because the curvature foreshortening is ignored in the intrinsic setup for mathematical convenience. It is worth further exploring which one of the two differences causes the drastic change in the uniqueness property.

- The intrinsic setup shares strong connections to the differential geometry analysis by Kunsberg and Zucker [42, 43], but they are still essentially different because of the approximation in our imaging model that ignores the curvature foreshortening effect. One advantage of Kunsberg and Zucker's work is that the curvature information can be inferred from shading in a lighting-invariant way, and showing direct and concrete connection to their work will help us understand how to apply our theory and algorithms to perform shape inference without the knowledge of lighting, or at least in a more robust way with respect to lighting. To do so, we will need to derive intrinsic uniqueness properties without ignoring curvature foreshortening, which potentially needs new or different mathematical tools.

In Chapter 4, we introduce a multi-scale framework for physics-based inference that uses a dense, overlapping set of image regions with local models, and show that when the regions are organized into a spatial hierarchy, inference can be done in an efficient and parallel way. We demonstrate the superior performance of framework on the binocular stereo application with respect to other state-of-the-art algorithms, but there are still a few directions that need to be explored and can make the framework useful for more general visual inference tasks:

- The current framework still depends on a well-performing domain-specific data cost, such as semi-global matching (SGM) [30] in binocular stereo. One important direction of future research is to alleviate this dependency and make the framework more robust to the input data cost. Section 4.6.1 presents potential ways of sampling a more generic data cost (which does not need to be quadratic or even smooth), but the computation complexity is still high. We have also explored the idea of sampling the parameter space adaptively instead of uniformly—that is first coarsely sample the parameter space, and then let the consensus output in each iteration guide the algorithm to perform more sampling at places that are more likely to yield a better local minima. Some preliminary experiments showed promising results, but the idea is still pre-mature and more work needs to be done to fully prove the efficacy of concept and to incorporate it into a robust algorithm.

- Another important direction of future research lies in applying the framework to problems involving estimating different physical properties of the same scene (such as material and shape), with different piecewise local models for each, when the aggregation regions of one property suggest, but do not determine, those of the other.

- Many properties of the consensus framework—multi-scale collaboration, implementation as a distributed architecture of computational units carrying out the same operations, coarse-to-fine evolution of the scene map, *etc*.—mimic behavior observed in biological systems [55]. It would be interesting to explore these links systematically—to investigate whether the framework, or some variation of it, can serve as a faithful model for biological processing; as well as whether insights from biology can be used to further improve the framework.

A

Appendix to Chapter 3

A.1 Proofs of Lemma 3.3

In this appendix, we provide a proof for Lemma 3.3 from Section 3.3.1. As a reminder, the lemma is defined in terms of matrices $A \in \mathbb{R}^{3\times 3}$ which are related to the coefficient vectors as:

$$A = \begin{bmatrix} -2a_1 & -a_3 & -a_4 \\ -a_3 & -2a_2 & -a_5 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.6}$$

The statement of the lemma itself is reproduced below.

Lemma 3.3 : *Let A and $\widetilde{A}$ correspond to two coefficient matrices of the form in* (3.6)*, and l and $\widetilde{l}$ to two lighting vectors. If,*

$$\frac{\bar{x}^T A^T l l^T A \bar{x}}{\bar{x}^T A^T A \bar{x}} = \frac{\bar{x}^T \widetilde{A}^T \widetilde{l}\widetilde{l}^T \widetilde{A} \bar{x}}{\bar{x}^T \widetilde{A}^T \widetilde{A} \bar{x}}, \forall \bar{x} \in \Omega, \tag{3.12}$$

$Rank(V_\Omega) = 15$, $Rank(A) \geq 2$, and $l^T A \bar{x} > 0, \forall \bar{x} \in \Omega$ (i.e., no point is in shadow), then

$$A^T l l^T A = \widetilde{A}^T \widetilde{l}\widetilde{l}^T \widetilde{A}, \quad A^T A = \widetilde{A}^T \widetilde{A}. \tag{3.13}$$

Moreover, if $Rank(A) = 2$, then $Rank(\widetilde{A}) = 2$ and both A and $\widetilde{A}$ have a common null space.

The expression in (3.12) equates two rational forms in x. To prove the lemma, we will show that the equality holds for all x if we have a sufficient number of non-degenerate locations in the patch. Then, we will show that the corresponding coefficients in the quadratic expressions in the numerator and denominator must be equal when they are of the form in (3.6) and the conditions of Lemma 3.3 are met, essentially ruling out the possibility of a common factor or scaling term. To this end, we introduce another lemma, with proof, and then present the proof of Lemma 3.3.

Lemma A.1. *Let $P, Q, \widetilde{P}, \widetilde{Q} \in \mathbb{R}^{3\times 3}$ be symmetric matrices. Then, $P = t\widetilde{P}$ and $Q = t\widetilde{Q}$, where $t \neq 0$ is a constant scalar, if*

$$\frac{\bar{x}^T P \bar{x}}{\bar{x}^T Q \bar{x}} = \frac{\bar{x}^T \widetilde{P} \bar{x}}{\bar{x}^T \widetilde{Q} \bar{x}}, \quad \forall \bar{x} \in \Omega, \tag{A.1}$$

$Rank(V_\Omega) = 15$, and,

Case 1: *All of the following conditions are satisfied:*

$$q_{11} \neq 0, \quad q_{22} \neq 0, \quad q_{33} \neq 0, \tag{A.2}$$

$$4(p_{11}q_{12} - p_{12}q_{11})(p_{22}q_{12} - p_{12}q_{22}) \; + \; (p_{11}q_{22} - p_{22}q_{11})^2 \neq 0, \tag{A.3}$$

$$4(p_{11}q_{13} - p_{13}q_{11})(p_{33}q_{13} - p_{13}q_{33}) \; + \; (p_{11}q_{33} - p_{33}q_{11})^2 \neq 0, \tag{A.4}$$

$$4(p_{22}q_{23} - p_{23}q_{22})(p_{33}q_{23} - p_{23}q_{33}) \; + \; (p_{22}q_{33} - p_{33}q_{22})^2 \neq 0. \tag{A.5}$$

Case 2: *All of the following conditions are satisfied:*

$$p_{j2}, p_{2j}, q_{j2}, q_{2j}, \widetilde{p}_{j2}, \widetilde{p}_{2j}, \widetilde{q}_{j2}, \widetilde{q}_{2j} = 0, \;\; \forall j \in \{1,2,3\}, \tag{A.6}$$

$$q_{11} \neq 0, \qquad q_{33} \neq 0, \tag{A.7}$$

$$4(p_{11}q_{13} - p_{13}q_{11})(p_{33}q_{13} - p_{13}q_{33}) \; + \; (p_{11}q_{33} - p_{33}q_{11})^2 \neq 0. \tag{A.8}$$

Proof of Lemma A.1: We re-write (A.1) as

$$\left(\bar{x}_i^T P \bar{x}_i\right) \cdot \left(\bar{x}_i^T \widetilde{Q} \bar{x}_i\right) = \left(\bar{x}_i^T \widetilde{P} \bar{x}_i\right) \cdot \left(\bar{x}_i^T Q \bar{x}_i\right), \tag{A.9}$$

and note that this is fourth-order polynomial equation in $\bar{x}_i$. Combining these equations $\forall \bar{x}_i \in \Omega$, we have

$$V_\Omega C_{(P,Q,\widetilde{P},\widetilde{Q})} = 0, \tag{A.10}$$

where $C_{(P,Q,\widetilde{P},\widetilde{Q})} \in \mathbb{R}^{15}$ are the coefficients of the polynomial, and are of the form $(p_{ij}\widetilde{q}_{kl} - \widetilde{p}_{ij}q_{kl})$. Since V_Ω is rank 15, the above equation implies that $C_{(P,Q,\widetilde{P},\widetilde{Q})} = 0$. We now consider different sets of coefficients to prove the lemma.

Case 1. First look at the coefficients of $x^4, y^4, x^3y, xy^3, x^2y^2$:

$$x^4: \quad p_{11}\widetilde{q}_{11} = \widetilde{p}_{11}q_{11} \tag{A.11}$$

$$y^4: \quad p_{22}\widetilde{q}_{22} = \widetilde{p}_{22}q_{22} \tag{A.12}$$

$$x^3y: \quad p_{11}\widetilde{q}_{12} + p_{12}\widetilde{q}_{11} = \widetilde{p}_{11}q_{12} + \widetilde{p}_{12}q_{11} \tag{A.13}$$

$$xy^3: \quad p_{22}\widetilde{q}_{12} + p_{12}\widetilde{q}_{22} = \widetilde{p}_{22}q_{12} + \widetilde{p}_{12}q_{22} \tag{A.14}$$

$$x^2y^2: \quad p_{11}\widetilde{q}_{22} + 4p_{12}\widetilde{q}_{12} + p_{22}\widetilde{q}_{11} = \widetilde{p}_{11}q_{22} + 4\widetilde{p}_{12}q_{12} + \widetilde{p}_{22}q_{11} \tag{A.15}$$

Since $q_{11} \neq 0, q_{22} \neq 0$, we can define $t = \widetilde{q}_{11}/q_{11}$ and $s = \widetilde{q}_{22}/q_{22}$. Then (A.11) and (A.12) gives us

$$\widetilde{q}_{11} = q_{11}t, \quad \widetilde{p}_{11} = p_{11}t, \quad \widetilde{q}_{22} = q_{22}s, \quad \widetilde{p}_{22} = p_{22}s. \tag{A.16}$$

Substitute into (A.13), (A.14), and (A.15), we have

$$\begin{aligned} -q_{11}\widetilde{p}_{12} + p_{11}\widetilde{q}_{12} &= (p_{11}q_{12} - p_{12}q_{11})t, \\ (p_{12}q_{22} - p_{22}q_{12})s - q_{22}\widetilde{p}_{12} + p_{22}\widetilde{q}_{12} &= 0, \\ (p_{11}q_{22} - p_{22}q_{11})s - 4q_{12}\widetilde{p}_{12} + 4p_{12}\widetilde{q}_{12} &= (p_{11}q_{22} - p_{22}q_{11})t. \end{aligned} \tag{A.17}$$

This can be thought of as a linear system of equations on $(s, \widetilde{p}_{12}, \widetilde{q}_{12})$, with one obvious solution $(t, p_{12}t, q_{12}t)$. This solution will be unique when the corresponding coefficient matrix is non-singular,

$$\det \begin{vmatrix} 0 & -q_{11} & p_{11} \\ (p_{12}q_{22} - p_{22}q_{12}) & -q_{22} & p_{22} \\ (p_{11}q_{22} - p_{22}q_{11}) & -4q_{12} & 4p_{12} \end{vmatrix} \neq 0. \tag{A.18}$$

Expanding this gives us (A.3), and therefore we have

$$[\widetilde{p}_{11}, \widetilde{p}_{12}, \widetilde{p}_{22}\widetilde{q}_{11}, \widetilde{q}_{12}, \widetilde{q}_{22}] = t[p_{11}, p_{12}, p_{22}, q_{11}, q_{12}, q_{22}]. \tag{A.19}$$

This approach can be used to show the same relationship for other terms in $P, Q, \widetilde{P}, \widetilde{Q}$. Specifically, the coefficients of $\{x^4, x^3, x^2, x, 1\}$ give us that $[q_{11}, q_{13}, q_{33}, p_{11}, p_{13}, p_{33}]$ and $[\widetilde{q}_{11}, \widetilde{q}_{13}, \widetilde{q}_{33}, \widetilde{p}_{11}, \widetilde{p}_{13}, \widetilde{p}_{33}]$ are proportional, and since they are linked by q_{11} and $\widetilde{q}_{11}$, the constant of proportionality must also be t. Similarly, looking at the coefficients of $\{y^4, y^3, y^2, y, 1\}$ gives us that $[q_{22}, q_{23}, q_{33}, p_{33}, p_{23}, p_{33}]$ and $[\widetilde{q}_{22}, \widetilde{q}_{23}, \widetilde{q}_{33}, \widetilde{p}_{33}, \widetilde{p}_{23}, \widetilde{p}_{33}]$ are proportional, with q_{22} and $\widetilde{q}_{22}$ linking the proportionality constant to t.

Case 2. For this case, we need to only look at the coefficients of $\{x^4, x^3, x^2, x, 1\}$, which gives us $[q_{11}, q_{13}, q_{33}, p_{11}, p_{13}, p_{33}] = t[\widetilde{q}_{11}, \widetilde{q}_{13}, \widetilde{q}_{33}, \widetilde{p}_{11}, \widetilde{p}_{13}, \widetilde{p}_{33}]$. □

Proof of Lemma 3.3: Without loss of generality, we rotate and translate the co-ordinate system so that $a_3 = 0$, and $(0, 0) \in \Omega$, and define $P = A^T l l^T A$, $Q = A^T A$, $\widetilde{P} = \widetilde{A}^T \widetilde{l}\, \widetilde{l}^T \widetilde{A}$ and $\widetilde{Q} = \widetilde{A}^T \widetilde{A}$. We consider two cases corresponding to the rank of A.

Case 1. Rank$(A) = 3$: We apply case 1 of Lemma A.1 by showing that the conditions (A.2)-(A.5) hold:

1. Since A is invertible, we have $q_{11} = 4a_1^2 \neq 0$ and $q_{22} = 4a_2^2 \neq 0$. Also, $q_{33} = a_4^2 + a_5^2 + 1 \neq 0$, and therefore, (A.2) is satisfied.

2. For (A.3) to be satisfied, we need

$$256 a_1^4 a_2^4 \left(l_x^2 + l_y^2\right)^2 \neq 0, \tag{A.20}$$

where $l = [l_x, l_y, l_z]$. Since A is invertible, $a_1 \neq 0$ and $a_2 \neq 0$. Note that (A.3) is violated if $l_x = l_y = 0$ and $\widetilde{l}_x = \widetilde{l}_y = 0$ (if not the latter, we can switch $\{a, l\}$, and $\{\widetilde{a}, \widetilde{l}\}$), but in that case, it is easy to see that

$$A^T l l^T A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & l_z^2 \end{bmatrix}, \; \widetilde{A}^T \widetilde{l}\,\widetilde{l}^T \widetilde{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \widetilde{l}_z^2 \end{bmatrix}, \tag{A.21}$$

which in turn implies $A^T l l^T A = t\widetilde{A}^T \widetilde{l}\,\widetilde{l}^T \widetilde{A}$, with $t = l_z^2/\widetilde{l}_z^2$.

3. For (A.4)-(A.5) to be satisfied, we need

$$16a_1^4 \left((a_5^2 + 1)l_x^2 + (l_z - a_5 l_y)^2\right)^2 \neq 0, \tag{A.22}$$

$$16a_2^4 \left((a_4^2 + 1)l_y^2 + (l_z - a_4 l_x)^2\right)^2 \neq 0. \tag{A.23}$$

Since $a_1, a_2 \neq 0$, these conditions will be violated when $l_x = 0$ and $l_z - a_5 l_y = 0$; or $l_y = 0$ and $l_z - a_4 l_x = 0$, respectively. But these cases can be ruled out, since they result in the point $(0, 0)$ being in shadow.

Therefore, from case 1 of Lemma A.1 we have that

$$A^T l l^T A = t\widetilde{A}^T \widetilde{l}\,\widetilde{l}^T \widetilde{A}, \;\; A^T A = t\widetilde{A}^T \widetilde{A}. \tag{A.24}$$

To show $t = 1$, we first look at the top-left 2×2 block of the matrix

$$\begin{bmatrix} 4a_1^2 & \\ & 4a_2^2 \end{bmatrix} = t \begin{bmatrix} 4\widetilde{a}_1^2 + \widetilde{a}_3^2 & 2\widetilde{a}_3(\widetilde{a}_1 + \widetilde{a}_2) \\ 2\widetilde{a}_3(\widetilde{a}_1 + \widetilde{a}_2) & 4\widetilde{a}_2^2 + \widetilde{a}_3^2 \end{bmatrix}, \tag{A.25}$$

which implies

$$a_1 = p_1\sqrt{t}\widetilde{a}_1,\ a_2 = p_2\sqrt{t}\widetilde{a}_2,\ \widetilde{a}_3 = 0, \tag{A.26}$$

with $p_1 = \pm 1, p_2 = \pm 1$. Next compare the $(1,3)$ and $(2,3)$ entry of the matrices (A.24), we have

$$\begin{cases} 2a_1a_4 = & 2t\widetilde{a}_1\widetilde{a}_4 \\ 2a_2a_5 = & 2t\widetilde{a}_2\widetilde{a}_5 \end{cases}, \quad \Rightarrow \quad \begin{cases} a_4 = & p_1\sqrt{t}\widetilde{a}_4 \\ a_5 = & p_2\sqrt{t}\widetilde{a}_5 \end{cases}. \tag{A.27}$$

Finally, look at the $(3,3)$ entry of the matrices in (A.24)

$$1 + a_4^2 + a_5^2 = t(1 + \widetilde{a}_4^2 + \widetilde{a}_5^2), \quad \Rightarrow \quad t = 1. \tag{A.28}$$

Case 2. Rank$(A) = 2$: Again, without loss of generality, we assume that the rank deficiency in A is caused by a_2 being equal to 0. Before we can apply case 2 of Lemma A.1, we need to show that there is no possible solution for $(\widetilde{a}, \widetilde{l})$ where $\widetilde{a}_2 \neq 0$, or $\widetilde{a}_3 \neq 0$. To do so, we look at the expression for $I_{\bar{x}}$ in terms of a and l:

$$I_{\bar{x}} = \frac{-(2a_1x + a_4)l_x - a_5l_y + l_z}{\sqrt{(2a_1x + a_4)^2 + a_5^2 + 1}}. \tag{A.29}$$

Note that the intensity here is independent of the coordinate y. Since $(\widetilde{a}, \widetilde{l})$ produce the same set of intensities, they too must be independent of y, which implies that $\widetilde{a}_2 = \widetilde{a}_3 = 0$.

We can then simply apply case 2 of Lemma A.1, using the same approach as in case 1 above, where (A.6),(A.7) are directly satisfied by the constraints on a and that $\widetilde{a}_2, \widetilde{a}_3 = 0$, and (A.8) is satisfied by $a_1 \neq 0$, and the constraint that the point $(0,0)$ not be in shadow. □

A.2 Proofs of Theorem 3.6

Theorem 3.6. *Given intensities $I(x, y)$ in an image patch Ω, generated by approximate imaging model in Definition 3.5 from a patch/lighting pair $(\mathcal{P}(\theta, a_1, a_2, a_3), l)$, then generically for any given lighting $\widetilde{l}$, there are at most four intrinsic quadratic patches $\mathcal{P}(\widetilde{\theta}, \widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3)$ that can exactly explain the image.*

We first write the Lambertian rendering equation (3.39) in a matrix form, and then elaborate the implication of approximate imaging model in Definition 3.5 to uniqueness properties of intrinsic quadratic patch. Define a *symmetric* matrix

$$\boldsymbol{A} = \begin{bmatrix} -2a_1 & -a_3 \\ -a_3 & -2a_2 \end{bmatrix}, \tag{A.30}$$

we can re-write the un-normalized normal vector of the patch $\mathbf{n}_0 = [n_s, n_t, 1]^T$ in matrix form

$$\begin{bmatrix} n_s \\ n_t \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} s \\ t \end{bmatrix}. \tag{A.31}$$

For notational simplicity, we further define

$$l_{[\boldsymbol{F}]} = \begin{bmatrix} l_s \\ 0 \\ l_n \end{bmatrix}, \quad \underline{\mathbf{s}} = \begin{bmatrix} s \\ t \\ 1 \end{bmatrix}, \quad \underline{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & \\ & 1 \end{bmatrix}, \tag{A.32}$$

then the Lambertian intensity in (3.39) can be written in a the following matrix form

$$I(s, t) = \frac{n_s l_s + l_n}{\sqrt{n_s^2 + n_t^2 + 1}} = \frac{l_{[\boldsymbol{F}]}^T \underline{\boldsymbol{A}}\, \underline{\mathbf{s}}}{\sqrt{\underline{\mathbf{s}}^T \underline{\boldsymbol{A}}^T \underline{\boldsymbol{A}}\, \underline{\mathbf{s}}}}. \tag{A.33}$$

Definition 3.5 implies a linear relationship between image coordinate (x, y) and the intrinsic patch coordinate (s, t) with a 2×2 matrix $\boldsymbol{P}$, and therefore

$$I(x, y) = I(\boldsymbol{P}^{-1}(s, t)) = \frac{l_{[\boldsymbol{F}]}^T \underline{\boldsymbol{A}}\, \underline{\boldsymbol{P}}^{-1} \underline{\mathbf{x}}}{\sqrt{\underline{\mathbf{x}}^T \underline{\boldsymbol{P}}^{-T} \underline{\boldsymbol{A}}^T \underline{\boldsymbol{A}}\, \underline{\boldsymbol{P}}^{-1} \underline{\mathbf{x}}}} = \frac{l_{[\boldsymbol{F}]}^T \underline{\boldsymbol{W}}\, \underline{\mathbf{x}}}{\sqrt{\underline{\mathbf{x}}^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{W}}\, \underline{\mathbf{x}}}}, \tag{A.34}$$

where

$$\underline{\mathbf{x}} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad \underline{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{P} & \\ & 1 \end{bmatrix}, \quad \underline{\boldsymbol{W}} = \underline{\boldsymbol{A}}\, \underline{\boldsymbol{P}}^{-1}. \tag{A.35}$$

Next we state and prove two lemmas, whose results will lead to our final proof to Theorem 3.6.

Lemma A.2. *All entries in $\boldsymbol{P}$ matrix are linear combination of* $\sin\theta$ *and* $\cos\theta$.

Proof. First note that matrix $\boldsymbol{P}$ is the upper-left 2×2 block of frame matrix $\boldsymbol{F} = [\mathbf{f}_s, \mathbf{f}_t, \mathbf{n}_0]$. Therefore we only need to show that all entries in $\boldsymbol{F}$ matrix are linear combination of $\sin\theta$ and $\cos\theta$. We choose a local reference frame $\boldsymbol{F}^{(0)}$ with $\theta = 0$, and any other local frame $\boldsymbol{F}^{(\theta)}$ with parameter θ can be obtained by a rotation

$$\boldsymbol{F}^{(\theta)} = \boldsymbol{R}^{(l,\theta)} \boldsymbol{F}^{(0)}, \tag{A.36}$$

where $\boldsymbol{R}^{(l,\theta)}$ is a rotation matrix that rotates the frame across axis l for an angle θ. More specifically,

$$\boldsymbol{R}^{(l,\theta)} = \begin{bmatrix} \cos\theta + l_x^2(1-\cos\theta) & l_x l_y(1-\cos\theta) - l_z\sin\theta & l_x l_z(1-\cos\theta) + l_y\sin\theta \\ l_x l_y(1-\cos\theta) + l_z\sin\theta & \cos\theta + l_y^2(1-\cos\theta) & l_y l_z(1-\cos\theta) - l_x\sin\theta \\ l_z l_x(1-\cos\theta) - l_y\sin\theta & l_z l_y(1-\cos\theta) + l_x\sin\theta & \cos\theta + l_z^2(1-\cos\theta) \end{bmatrix}, \tag{A.37}$$

and note that each entry in the matrix is a linear combination of $\sin\theta$ and/or $\cos\theta$. Finally, we can

write the orthographical projection matrix $\boldsymbol{P}$ as

$$\boldsymbol{P}^{(\theta)} = \boldsymbol{R}^{(l,\theta)}_{1:2,1:2}\boldsymbol{F}^{(0)}_{1:2,1:2} + \boldsymbol{R}^{(l,\theta)}_{1:2,3}\boldsymbol{F}^{(0)}_{3,1:2}, \tag{A.38}$$

whose entries are also linear combinations of $\sin\theta$ and $\cos\theta$. □

Lemma A.3. *Let*

$$\underline{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{W} & \\ & 1 \end{bmatrix}, \quad \underline{\widetilde{\boldsymbol{W}}} = \begin{bmatrix} \widetilde{\boldsymbol{W}} & \\ & 1 \end{bmatrix} \tag{A.39}$$

correspond to two coefficient matrices of the form in (A.35), and

$$l = (l_s, 0, l_n)^T, \quad \widetilde{\boldsymbol{\ell}} = \left(\widetilde{l}_s, 0, \widetilde{l}_n\right)^T, \qquad l_s \geq 0, \widetilde{l}_s \geq 0 \tag{A.40}$$

two lighting vector in each of the two local frames. If

$$\frac{l^T \underline{\boldsymbol{W}}\, \underline{\mathbf{x}}}{\sqrt{\underline{\mathbf{x}}^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{W}}\, \underline{\mathbf{x}}}} = \frac{\widetilde{\boldsymbol{\ell}}^T \underline{\widetilde{\boldsymbol{W}}}\, \underline{\mathbf{x}}}{\sqrt{\underline{\mathbf{x}}^T \underline{\widetilde{\boldsymbol{W}}}^T \underline{\widetilde{\boldsymbol{W}}}\, \underline{\mathbf{x}}}} \tag{A.41}$$

hold for a non-degenerate set of $\underline{\mathbf{x}}$*, then generically, we have*

$$l = \widetilde{\boldsymbol{\ell}}, \qquad \underline{\widetilde{\boldsymbol{W}}} = \begin{bmatrix} 1 & & \\ & \pm 1 & \\ & & 1 \end{bmatrix} \underline{\boldsymbol{W}}. \tag{A.42}$$

Proof. Since $\underline{\boldsymbol{W}}$ and $\underline{\widetilde{\boldsymbol{W}}}$ are affine matrices, and *in the generic setup*, assuming $\boldsymbol{W}$ matrix is full-rank, we can apply the Lemma 3.3, which implies

$$\underline{\boldsymbol{W}}^T l l^T \underline{\boldsymbol{W}} = \underline{\widetilde{\boldsymbol{W}}}^T \widetilde{\boldsymbol{\ell}} \widetilde{\boldsymbol{\ell}}^T \underline{\widetilde{\boldsymbol{W}}}, \qquad \underline{\boldsymbol{W}}^T \underline{\boldsymbol{W}} = \underline{\widetilde{\boldsymbol{W}}}^T \underline{\widetilde{\boldsymbol{W}}}. \tag{A.43}$$

Substituting in the first equation with the special form of $\underline{\boldsymbol{W}}$, $\underline{\widetilde{\boldsymbol{W}}}$ in (A.39) and $l, \widetilde{\boldsymbol{\ell}}$ in (A.40), we have

$$\begin{bmatrix} l_s w_{11} \\ l_s w_{12} \\ l_n \end{bmatrix} \begin{bmatrix} l_s w_{11} & l_s w_{12} & l_n \end{bmatrix} = \begin{bmatrix} \widetilde{l}_s \widetilde{w}_{11} \\ \widetilde{l}_s \widetilde{w}_{12} \\ \widetilde{l}_n \end{bmatrix} \begin{bmatrix} \widetilde{l}_s \widetilde{w}_{11} & \widetilde{l}_s \widetilde{w}_{12} & \widetilde{l}_n \end{bmatrix}. \tag{A.44}$$

Comparing the (3,3)-entry, with the knowledge that the central pixel is not in shadow, we have $l_n = \widetilde{l}_n$. Further comparing the (1,3)- and (2,3)-entry gives us

$$l_s w_{11} = \widetilde{l}_s \widetilde{w}_{11}, \tag{A.45}$$

$$l_s w_{12} = \widetilde{l}_s \widetilde{w}_{12}. \tag{A.46}$$

Generically, assume $\widetilde{l}_s \neq 0$, and define $t = l_s / \widetilde{l}_s$, and we have $\widetilde{w}_{11} = t w_{11}, \widetilde{w}_{12} = t w_{12}$.

Now look at the second equation in (A.43), and substitute in the special form of $\underline{\boldsymbol{W}}$, $\underline{\widetilde{\boldsymbol{W}}}$, we have

$$w_{11}^2 + w_{21}^2 = \widetilde{w}_{11}^2 + \widetilde{w} w_{21}^2 \qquad \Longrightarrow \qquad \widetilde{w}_{21}^2 = (1 - t^2) w_{11}^2 + w_{21}^2 \tag{A.47}$$

$$w_{11} w_{12} + w_{21} w_{22} = \widetilde{w}_{11} \widetilde{w}_{12} + \widetilde{w}_{21} \widetilde{w}_{22} \qquad \Longrightarrow \qquad \widetilde{w}_{21} \widetilde{w}_{22} = (1 - t^2) w_{11} w_{12} + w_{21} w_{22} \tag{A.48}$$

$$w_{12}^2 + w_{22}^2 = \widetilde{w}_{12}^2 + \widetilde{w}_{22}^2 \qquad \Longrightarrow \qquad \widetilde{w}_{22}^2 = (1 - t^2) w_{12}^2 + w_{22}^2 \tag{A.49}$$

Multiplying (A.47) and (A.49) and comparing with the square of (A.48), we have

$$(1 - t^2)(w_{11}^2 w_{22}^2 + w_{21}^2 w_{12}^2) = 2(1 - t^2) w_{11} w_{12} w_{21} w_{22}. \tag{A.50}$$

Generically, assume $w_{11} w_{22} \neq w_{12} w_{21}$, we have $w_{11}^2 w_{22}^2 + w_{21}^2 w_{12}^2 > 2 w_{11} w_{12} w_{21} w_{22}$, and therefore $t^2 = 1$. Since $l_s > 0$ and $\widetilde{l}_s > 0$, we have $t = 1$, *i.e.* $l_s = \widetilde{w} l_s$.

Finally, going back to (A.47)-(A.49), we have

$$\begin{bmatrix} w_{21} \\ w_{22} \end{bmatrix} = \pm \begin{bmatrix} \widetilde{w}_{21} \\ \widetilde{w}_{22} \end{bmatrix},$$

sign cannot be resolved from the information we have. □

Proof of Theorem 3.6: Suppose there are two intrinsic quadratic patches / lighting pair $\{\mathcal{P}(\theta, a_1, a_2, a_3), l\}$ and $\{\mathcal{P}\left(\widetilde{\theta}, \widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3\right), \widetilde{\boldsymbol{\ell}}\}$ that generate the same observed image $I(x, y)$. According to Lemma A.3, we have

$$\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{P}}^{-1} = \begin{bmatrix} 1 & \\ & \pm 1 \end{bmatrix} \boldsymbol{A}\boldsymbol{P}^{-1} \qquad \Longrightarrow \qquad \widetilde{\boldsymbol{A}} = \begin{bmatrix} 1 & \\ & \pm 1 \end{bmatrix} \boldsymbol{A}\boldsymbol{P}^{-1}\widetilde{\boldsymbol{P}}. \tag{A.51}$$

Note that $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$ are a symmetric matrices with parameter a_1, a_2, a_3 and $\widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3$, and that $\boldsymbol{P}$ and $\widetilde{\boldsymbol{P}}$ is a function of θ and $\widetilde{\theta}$ as shown in (A.38).

Now we take patch $\mathcal{P}(\theta, a_1, a_2, a_3)$ as ground truth, and consider the problem that "what set of possible $\left(\widetilde{\theta}, \widetilde{a}_1, \widetilde{a}_2, \widetilde{a}_3\right)$ could satisfy (A.51)". The only constraint here is that $\widetilde{\boldsymbol{A}}$ needs to be a symmetric matrix. Taking difference of its $(1, 2)$-entry and $(2, 1)$-entry and equating it to zero will result in a linear equation on the unknown variable $\widetilde{\theta}$, and as proved in Lemma A.2, this equation must be linear with respect to $\sin\theta$ and $\cos\theta$, which is

$$\alpha_i \sin\theta + \beta_i \cos\theta + \gamma_i = 0, \tag{A.52}$$

with $i = 1, 2$ denoting the plus or minus sign of Equation (A.51). *Generically*, there are exactly two solutions to each of the equation (A.52), and therefore four in total when counting the sign flip. □

References

[1] Ackermann, J., Ritz, M., Stork, A., & Goesele, M. (2010). Removing the example from photometric stereo by example. In *Proc. Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments*.

[2] Barron, J. T. & Malik, J. (2013). Shape, illumination, and reflectance from shading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[3] Birchfield, S. & Tomasi, C. (1998). A pixel dissimilarity measure that is insensitive to image sampling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(4), 401–406.

[4] Birchfield, S. & Tomasi, C. (1999). Multiway cut for stereo and motion with slanted surfaces. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[5] Brady, M. & Legge, G. (2009). Camera calibration for natural image studies and vision research. *Journal of the Optical Society of America A*, 26(1), 30–42.

[6] Bredies, K., Kunisch, K., & Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*.

[7] Bruhn, A., Weickert, J., Kohlberger, T., & Schnörr, C. (2006). A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. *International Journal of Computer Vision*.

[8] Chakrabarti, A., Scharstein, D., & Zickler, T. (2009). An empirical camera model for internet color vision. In *British Machine Vision Conference*.

[9] Chakrabarti, A., Xiong, Y., Sun, B., Darrell, T., Scharstein, D., Zickler, T., & Saenko, K. (2014). Modeling radiometric uncertainty for vision with tone-mapped color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11), 2185–2198.

[10] Chen, X., Li, F., Yang, J., & Yu, J. (2012). A theoretical analysis of camera response functions in image deblurring. *European Conference on Computer Vision*.

[11] Cole, F., Isola, P., Freeman, W. T., Durand, F., & Adelson, E. H. (2012). Shapecollage: occlusion-aware, example-based shape interpretation. In *European Conference on Computer Vision*.

[12] dcraw (Last accessed: Janurary 10, 2011.). Decoding raw digital photos in linux. http://www.cybercom.net/ ~dcoffin/dcraw/.

[13] Debevec, P. & Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH* (pp. 369–378).

[14] Douglas, J. & Rachford, H. H. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, (pp. 421–439).

[15] Durou, J. D., Falcone, M., & Sagona, M. (2008). Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*.

[16] Ecker, A. & Jepson, A. D. (2010). Polynomial shape from shading. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[17] Einecke, N. & Eggert, J. (2014). Block-matching stereo with relaxed fronto-parallel assumption. In *Proc. IEEE Intelligent Vehicles Symposium*.

[18] Farid, H. (2002). Blind inverse gamma correction. *Image Processing, IEEE Transactions on*, 10(10), 1428–1433.

[19] Frankot, R. T. & Chellappa, R. (1988). A method for enforcing integrability in shape from shading algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 10(4), 439–451.

[20] Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision*.

[21] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 3354–3361).

[22] Grossberg, M. & Nayar, S. (2003). Determining the camera response from images: what is knowable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (pp. 1455–1467).

[23] Grossberg, M. D. & Nayar, S. K. (2004). Modeling the space of camera response functions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10), 1272–1282.

[24] Grundmann, M., McClanahan, C., Kang, S. B., & Essa, I. (2013). Post-processing approach for radiometric self-calibration of video. In *International Conference on Computational Photography* (pp. 1–9).

[25] Haber, T., Fuchs, C., Bekaer, P., Seidel, H., Goesele, M., & Lensch, H. (2009). Relighting objects from image collections. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 627–634).

[26] Haddon, J. & Forsyth, D. (1998). Shape representations from shading primitives. In *European Conference on Computer Vision*.

[27] Hasler, D. & Süsstrunk, S. (2004). Mapping colour in image stitching applications. *Journal of Visual Communication and Image Representation*, 15(1), 65–90.

[28] Hassner, T. & Basri, R. (2006). Example based 3D reconstruction from single 2D images. In *CVPR Workshop "Beyond patches"*.

[29] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[30] Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision and Pattern Recognition, IEEE Conference on*, volume 2 (pp. 807–814).

[31] Holm, J., Tastl, I., Hanlon, L., & Hubel, P. (2002). Color processing for digital photography. In P. Green & L. MacDonald (Eds.), *Colour Engineering: Achieving Device Independent Colour* (pp. 179–220). Wiley.

[32] Horn, B. K. & Brooks, M. J. (1986). The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*.

[33] Horn, B. K. P. & Brooks, M. J. (1989). *Shape from shading*. MIT press.

[34] Huang, X., Gao, J., Wang, L., & Yang, R. (2007). Examplar-based shape from shading. In *3-D Digital Imaging and Modeling*.

[35] Huggins, P. S., Chen, H. F., Belhumeur, P. N., & Zucker, S. W. (2001). Finding folds: On the appearance and identification of occlusion. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[36] Johnson, M. K. & Adelson, E. H. (2011). Shape estimation in natural illumination. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[37] Kanade, T. & Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: Theory and experiment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(9), 920–932.

[38] Kim, M. & Kautz, J. (2008). Characterization for high dynamic range imaging. *Computer Graphics Forum (Proc. EGSR)*, 27(2), 691–697.

[39] Kim, S., Lin, H., Lu, Z., Süsstrunk, S., Lin, S., & Brown, M. (2012). A new in-camera imaging model for color computer vision and its application. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[40] Korman, S., Reichman, D., Tsur, G., & Avidan, S. (2013). Fast-match: Fast affine template matching. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 2331–2338).

[41] Krähenbühl, P. & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Neural Information Processing Systems*.

[42] Kunsberg, B. & Zucker, S. W. (2013). Characterizing ambiguity in light source invariant shape from shading. *arXiv:1306.5480v1*.

[43] Kunsberg, B. & Zucker, S. W. (2014). How shading constrains surface patches without knowledge of light sources. *SIAM Journal on Imaging Sciences*, 7(2), 641–668.

[44] Kuschk, G. & Cremers, D. (2013). Fast and accurate large-scale stereo reconstruction using variational methods. In *Proc. ICCV Workshops*.

[45] Kuthirummal, S., Agarwala, A., Goldman, D., & Nayar, S. (2008). Priors for large photo collections and what they reveal about cameras. In *European Conference on Computer Vision*.

[46] Lalonde, J., Narasimhan, S., & Efros, A. (2010). What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 88(1), 24–51.

[47] Lempitsky, V., Roth, S., & Rother, C. (2008). Fusionflow: Discrete-continuous optimization for optical flow estimation. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 1–8).

[48] Lempitsky, V., Vedaldi, A., & Zisserman, A. (2011). Pylon model for semantic segmentation. In *Neural Information Processing Systems*.

[49] Lin, H., Kim, S. J., Susstrunk, S., & Brown, M. (2011). Revisiting radiometric calibration for color computer vision. In *International Conference on Computer Vision*.

[50] Lin, S., Gu, J., Yamazaki, S., & Shum, H.-Y. (2004). Radiometric calibration from a single image. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[51] Mann, S. & Picard, R. (1995). Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proc. IS&T Annual Conference* (pp. 422–428).

[52] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*.

[53] Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*.

[54] McHutchon, A. & Rasmussen, C. E. (2011). Gaussian process training with input noise. In *Neural Information Processing Systems*.

[55] Menz, M. D. & Freeman, R. D. (2004). Functional connectivity of disparity-tuned neurons in the visual cortex. *Journal of Neurophysiology*.

[56] Mitsunaga, T. & Nayar, S. (1999). Radiometric self calibration. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[57] Oliensis, J. (1991a). Shape from shading as a partially well-constrained problem. *CVGIP: Image Understanding*.

[58] Oliensis, J. (1991b). Uniqueness in shape from shading. *International Journal of Computer Vision*, 6(2), 75–104.

[59] Owens, T., Saenko, K., Chakrabarti, A., Xiong, Y., Zickler, T., & Darrell, T. (2011). Learning object color models from multi-view constraints. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 169–176).

[60] Pal, C., Szeliski, R., Uyttendaele, M., & Jojic, N. (2004). Probability models for high dynamic range imaging. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[61] Pentland, A. P. (1984). Local shading analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[62] Prados, E. & Faugeras, O. (2004). Unifying approaches and removing unrealistic assumptions in shape from shading: Mathematics can help. In *International Conference on Computer Vision* (pp. 141–154).

[63] Prados, E. & Faugeras, O. (2005a). A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras. *International Journal of Computer Vision*.

[64] Prados, E. & Faugeras, O. (2005b). Shape from shading: a well-posed problem? In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[65] Ramanath, R., Snyder, W., Yoo, Y., & Drew, M. (2005). Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1), 34–43.

[66] Ranftl, R., Pock, T., & Bischof, H. (2013). Minimizing tgv-based variational models with non-convex data terms. In *Scale Space and Variational Methods in Computer Vision*.

[67] Rasmussen, C. E. & Ghahramani, Z. (2002). Infinite mixtures of gaussian process experts. In *Neural Information Processing Systems*.

[68] Rasmussen, C. R. & Williams, C. K. (2006). *Gaussian Process for Machine Learning*. MIT Press.

[69] Reinhard, E., Ward, G., Pattanaik, S., & Debevec, P. (2006). *High dynamic range imaging*. Elsevier.

[70] Rhemann, C., Hosni, A., Bleyer, M., Rother, C., & Gelautz, M. (2011). Fast cost-volume filtering for visual correspondence and beyond. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[71] Scharstein, D. & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3), 7–42.

[72] Shen, L. & Tan, P. (2009). Photometric stereo and weather estimation using internet images. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 1850–1857).

[73] Shi, B., Matsushita, Y., Wei, Y., Xu, C., & Tan, P. (2010). Self-calibrating photometric stereo. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[74] Slesareva, N., Bruhn, A., & Weickert, J. (2005). Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. In *Pattern Recognition*.

[75] Spangenberg, R., Langner, T., & Rojas, R. (2013). Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *Computer Analysis of Images and Patterns*.

[76] Spaulding, K. E., Gallagher, A. C., Gindele, E. B., & Ptucha, R. W. (2007). Constructing extended color gamut images from limited color gamut digital images. U.S. Patent No. 7,308,135.

[77] Sun, J., Zheng, N.-N., & Shum, H.-Y. (2003). Stereo matching using belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7), 787–800.

[78] Tai, Y., Chen, X., Kim, S., Li, F., Yang, J., Yu, J., Matsushita, Y., & Brown, M. (2013). Nonlinear camera response functions and image deblurring: Theoretical analysis and practice. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[79] Tan, P., Quan, L., & Zickler, T. (2011). The geometry of reflectance symmetries. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[80] Urtasun, R. & Darrell, T. (2008). Sparse probabilistic regression for activity-independent human pose inference. In *Computer Vision and Pattern Recognition, IEEE Conference on*.

[81] Vogel, C., Roth, S., & Schindler, K. (2013). Piecewise rigid scene flow. In *International Conference on Computer Vision*.

[82] Wagemans, J., Van Doorn, A. J., & Koenderink, J. J. (2010). The shading cue in context. *i-Perception*.

[83] Wei, D., Liu, C., & Freeman, W. (2014). A data-driven regularization model for stereo and flow. In *Proc. International Conferene on 3D Vision*.

[84] Woodford, O., Torr, P., Reid, I., & Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12), 2115–2128.

[85] Woodham, R. (1980). Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1), 139–144.

[86] Xiong, Y., Chakrabarti, A., Basri, R., Gortler, S. J., Jacobs, D. W., & Zickler, T. (2015). From shading to local shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[87] Yamaguchi, K., Hazan, T., McAllester, D., & Urtasun, R. (2012). Continuous markov random fields for robust stereo estimation. In *European Conference on Computer Vision*.

[88] Yamaguchi, K., McAllester, D., & Urtasun, R. (2013). Robust monocular epipolar flow estimation. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 1862–1869).

[89] Yamaguchi, K., McAllester, D., & Urtasun, R. (2014). Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*.

[90] Yoon, K.-J. & Kweon, I. S. (2006). Adaptive support-weight approach for correspondence search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[91] Zabih, R. & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision* (pp. 151–158). Springer.

[92] Žbontar, J. & LeCun, Y. (2014). Computing the stereo matching cost with a convolutional neural network. *arXiv:1409.4326*.

[93] Zhang, K., Fang, Y., Min, D., Sun, L., Yan, S. Y., Tian, Q., et al. (2014). Cross-scale cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 1590–1597).

[94] Zhang, R., Tsai, P., Cryer, J. E., & Shah, M. (1999). Shape from shading: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[95] Zhu, Q. & Shi, J. (2006). Shape from shading: Recognizing the mountains through a global view. In *Computer Vision and Pattern Recognition, IEEE Conference on*.